Vol.2014-DBS-160 No.11
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.11
2014/11/18

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

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

概要：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トピックを抽出・可視化する試作システムについて述べる．試作システムでは，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対し，関連するいくつかの英語の Wikipedia の記事をトピックとして付与する．そして，Wikipedia の記事をノード，同じテキストに付与された Wikipedia の記事の共起をエッジとするグラフを構築する．これにより，トピックの空間を統一した状態で，トピック間の関連を言語別に表現できる．また，Wikipedia の記事をクエリとして与えたとき，その記事に関連して話題となっているトピックを言語横断的に可視化できる．試作システムについて予備実験を行ない，試作システムがどの程度機能するかを検証する．また，実験結果からトピック抽出手法や可視化手法，評価方法のデザインについて考察する．

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，最近では，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 [*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．また，言語横断的に抽出したトピック情報について各言語における話題の度合やトピック間の関連性を可視化することで，言語や文化という視点からトピック情報を比較・調査することが可能になる．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する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・可視化し，それらのトピックを様々な視点から比較・調査ができるように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・可視化するにはいくつかの問題が存在する．まず，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である．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ためには，トピックの言語空間を統一する必要がある．また，トピック情報の可視化においては，抽出したトピック情報をどのような形で提示するかが問題となる．例えば，ランキングによる可視化では，それぞれの言語においてどのようなトピックが注目されているかを表現できるが，注目されているトピックについて言語間で差異があるのかどうかは表現できない．特定のトピックに関する言語間の差異を表現するためには，それぞれの言語において，そのトピックと同時に言及されやすいトピックにどのような違いがあるのか，といったトピック間の関連性を考慮し

[1] 大阪大学大学院情報科学研究科
[a)] nakamura.tatsuya@ist.osaka-u.ac.jp
[*1] 2014 年 10 月時点．ユーザ設定画面において確認 (Beta 版含む)．

たトピック情報の表現が必要になる．

本研究では，多言語なソーシャルメディアを対象として，ソーシャルメディア上で多くの人に言及され話題となっているトピック情報を言語横断的に抽出・可視化することを目的としたシステムを試作する．試作システムではまず，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エンティティリンキングを行い，テキスト中に明示的に出現するエンティティを Wikipedia の記事に紐付ける．このとき，英語の Wikipedia の記事をリンクさせることで，トピックの言語空間を英語に統一する．そして，同一のテキストに対して付与された記事の共起情報から，英語の Wikipedia の記事をノード，共起情報をエッジとするグラフを言語ごとに構築する．これにより，異なる言語のテキストから得られたトピックの言語空間を英語の Wikipedia の空間に統一した状態で，トピック間の関係を言語別に表現できる．また，Wikipedia の記事をクエリとして与えたとき，その記事に関連して話題となっているトピックを，言語別に構築したグラフを用いて言語横断的に可視化できる．

2. 関連研究

2.1 ソーシャルメディアを対象としたトピック抽出

ソーシャルメディアを対象としたトピック抽出に関する研究はこれまでにいくつか行われており，それらはトピックモデルを用いた研究 [18], [19] とトピック情報に関する木構造やグラフを用いた研究 [8], [20] に大別できる．Zhao ら [19] は，トピックモデルである LDA [2] を拡張し，短文の入力に対応した Twitter-LDA を提案している．また Zhao らは，Twitter-LDA を用いて抽出したトピックからキーワードを抽出する手法を提案している [18]．一般的に，LDA をベースとした手法によって抽出されたトピックは，そのトピックを構成するキーワードの集合によって表現されるため，トピックが表している概念を明示的に表現することは難しい．トピック間の関係を階層的に表現するトピックモデルも存在するが [1], [9]，上記のモデルと同様にトピックそのものは単語の集合として表現される．本研究では，トピックを一つの Wikipedia の記事により表現することで，トピックの内容を明示的に表現している．

Zhu らの研究 [20] では，Blog や Q&A，Twitter など複数のメディアのテキストから，単一の語句をトピックとした階層的なトピック構造を抽出する手法を提案している．この手法では，はじめに各メディアのテキストからキーワードを抽出する．次に，Wikipedia のカテゴリ情報や検索エンジンを用いたパターンマッチングによってキーワード間の階層関係を抽出し，キーワードをノード，キーワード間の階層関係をエッジとした木構造を構築する．この木構造に対して，特定のキーワードをクエリとして与えることで，そのキーワードをルートとしたトピック階層を出力する．また，抽出したトピック階層をリアルタイムに更新する手法についても提案している．このトピック抽出手法では，一つのトピックが一つの語句として明示的に表現される．しかし，トピック間の階層関係を抽出する際に外部の知識体系を用いているため，抽出されたトピックの階層が知識体系における語句の意味的な抽象度の度合を表す傾向にある．本研究では，話題となっているトピック情報を言語間で比較することを想定しているため，トピックの意味的な抽象度ではなく，実際に入力のテキスト集合中で言及されている度合を考慮したトピック構造を構築することを目指している．

Kang ら [8] は，ソーシャルメディアのテキストから，対話的に閲覧可能な階層的なタグクラウドを生成するシステム Vesta を提案している．また，抽出したトピック情報のタグクラウドによる可視化だけでなく，形式概念分析を用いてソーシャルメディアの膨大な量のデータから特徴的なキーワードを抽出・クラスタリングし，LDA を用いた手法より高速に処理する手法も提案している．生成された階層的なタグクラウドは，階層が深くなるほどより詳細なタグが表示される．実際のツイートから階層的なトピック構造を抽出している点で，本研究で抽出するトピック構造に近い特徴を持つと言える．

これらの手法は，単一言語のソーシャルメディアのテキストを対象とし，語句の共起情報やパターンマッチングによる語句間の関連抽出などを用いてトピックを抽出しているため，複数言語のテキストを対象としてトピック抽出を行うことは困難である．本研究では，任意の言語で記述されたテキストに対して，エンティティリンキングによりテキスト中のエンティティを英語の Wikipedia の記事にリンクし，トピックの言語空間を統一した後，それらの記事を対象としてトピック抽出を行うことで，多言語な文書集合からのトピック抽出を実現している．

2.2 多言語な文書集合を対象としたトピック抽出

多言語な文書集合を対象としたトピック抽出手法もいくつか提案されている．Ni ら [13] は，Wikipedia のような対訳関係が定義された多言語な文書集合からトピックを抽出する Multi-lingual LDA(ML-LDA) を提案し，文書分類のタスクにおいて抽出したトピックの有用性を示している．しかし，ML-LDA は異なる言語で記述されたテキスト間で対訳関係が定義されていることが必要であるため，ソーシャルメディアのテキストからトピック情報を直接抽出できない．一方，対訳関係が定義されていない文書集合を対象としたトピックモデル [3], [7] も提案されている．これらの手法についても，トピックモデルを用いた他の手法と同様に，抽出されたトピックは語句の集合として表現される．

トピックモデルのようにテキスト情報のみを用いてトピック情報を抽出する手法に対して，Wikipedia に代表さ

Vol.2014-DBS-160 No.11
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.11
2014/11/18

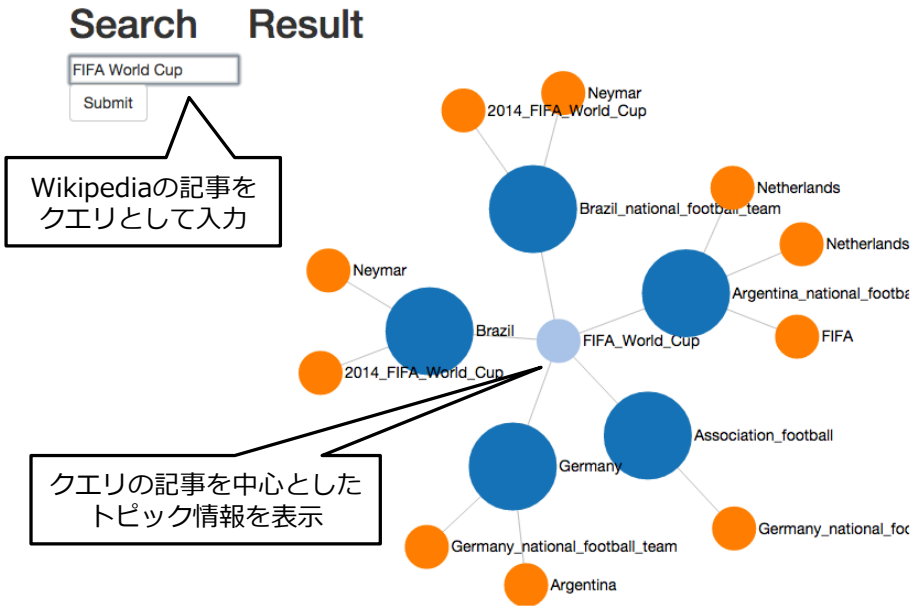

図 1 試作システムの使用例

れる多言語な知識体系を背景知識として用いることも考えられる．多言語な知識体系では，言語は異なるが同一のエンティティを表すエントリ間の対訳関係が定義されているため，エンティティリンキングによりテキストの内容に対応するエンティティを紐付けることができれば，エントリ間の対訳関係によって，異なる言語で記述されたテキスト間のトピック情報を比較できる．例えば Wikipedia では，言語間リンクと呼ばれるリンクにより，Wikipedia の記事 (エンティティ) 間の対訳関係が定義されている．中崎ら [21] は，言語間リンクを用いることで，予め用意したトピックに関する日本語と英語の特徴語を Wikipedia から収集し，その特徴語を用いてブログ記事をランキングすることで，日英ブログを対象とした言語横断的な検索を実現している．また，ブログ記事を人手により分析し，対象のトピックに関する言語間差異の対照分析を行っている．本研究では，多言語なソーシャルメディアのテキストを対象としたトピック抽出を目的としている点で，中崎らの研究と異なる．

3. 試作システムの概要

本研究で試作するシステムの機能は，任意の英語の Wikipedia の記事をクエリとして，その記事を中心として Twitter 上で話題になっている記事 (トピック) をグラフとして提示するというものである．図 1 は試作システムの使用例を示している．図 1 では，「FIFA World Cup」という英語の記事のクエリに対し，英語，スペイン語，日本語，アラビア語の 4 ヶ国語で共通して話題になっている記事が表示されている．また，クエリの記事からの距離が近い記事ほどクエリの記事との関連性やその記事の話題度が高い (実際にテキスト中で同時に言及されやすい) ことを表している．

試作システムを利用することにより，以下のような要求を満たすことができると考えられる．

- 関心のあるトピックに関する情報を入手する．ユーザは，関心のあるトピックを表す記事をクエリとして与えることで，その記事に関連して話題となっているトピック情報を検索できる．試作システムでは言語空間を統一した状態でトピック情報を保持しているため，言語を指定するだけで，指定した言語におけるトピックの情報を調べることができる．
- 関心のあるトピックについて，言語間での共通性や差異を調べる．各言語のトピック情報の言語空間が統一されているため，異なる言語間でのトピック情報を容易に比較できる．例えば，同一のクエリに対して各言語で共通して出現するトピックは，言語間で共通して話題になっているトピックであると考えられる．逆に，特定の言語でのみ出現するトピックは，その言語に特有のトピックを表していると考えられる．

4. 試作システムの処理の流れ

試作システムでは，対応する言語の集合を L として，任意の言語 $l \in L$ で記述されたテキストに対してエンティティリンキングにより英語の Wikipedia の記事を付与し，記事をノード，同じテキストに付与された記事の共起をエッジとするグラフを構築することで，言語横断的なトピック抽出を実現している (図 2)[*2]．ここで，任意の言語で記述さ

[*2] 付与する記事の言語として，最も記事数が多く，他の言語からの言語間リンクが多い英語を用いているが，本研究で提案するト

Vol.2014-DBS-160 No.11
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.11
2014/11/18

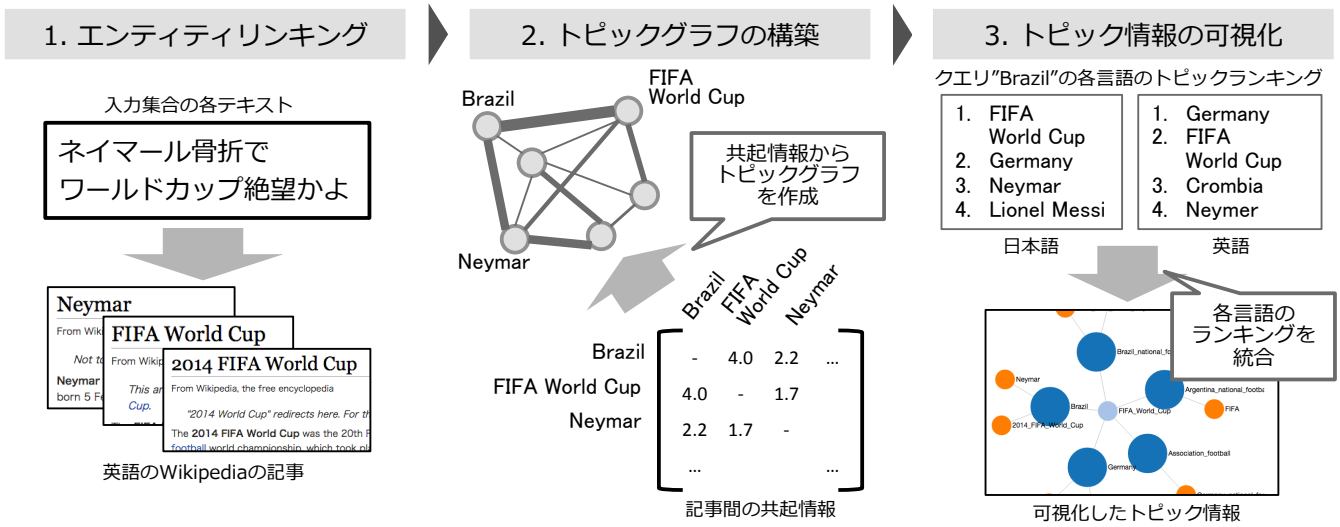

図 2 試作システムの処理の流れ

れた短いテキストに対して，どのように英語の Wikipedia の記事を付与するか，また，得られたグラフからどのようにしてクエリに対するトピック情報を抽出するかが技術的課題となる．以下の節では，図 2 に示す試作システムの各処理における課題とそれを解決するための要素技術について説明する．

4.1 TAGME によるエンティティリンキング

テキストの入力に対して，そのテキスト中に出現するエンティティを Wikipedia や Freebase などの知識体系のエントリに紐付けるタスクはエンティティリンキングと呼ばれる．特にソーシャルメディアのテキストような短いテキストを対象としたマイニングのタスクにおいては，エンティティリンキングによりテキスト自身が持つ情報量を増やすことが有効である [4], [11]．任意の言語で記述されたテキストに対して異なる言語のエンティティを付与する手法もいくつか存在する [10], [17]．しかし，これらの手法は明確に短文を入力として想定しておらず，また，識別器を用いているため学習データが必要であるといった問題がある．

そこで本研究では，短文を対象とし，かつ，識別器を使用しないエンティティリンキング手法である TAGME [4] を採用する *3．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から連想される記事 (エンティティ) 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．キーワードの抽出では，テキスト中に出現する Wikipedia のアンカーテキスト全て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 内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．
- $lp(a_1) \geq lp(a_2)$ の場合，a_1 と a_2 の両方をキーワードとして抽出する．

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，どの記事 $p_a \in Pg(a)$ を表しているかを次の式により算出する．

$$rel_a(p_a) = \sum_{b \in A \setminus \{a\}} \frac{\sum_{p_b \in Pg(b)} rel(p_b, p_a) \cdot Pr(p_b|b)}{|Pg(b)|} \quad (1)$$

ここで，$rel(p_b, p_a)$ は記事間関連度 [12] 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，式 (1) は記事 p_a が他のキーワードから連想される記事と互いに関連が強いほど高い値となる．そして，各キーワード a について式 (1) により得られるスコアの高い上位 $\epsilon\%$ の記事のうち，確率 $Pr(p_a|a)$ の最も高い記事 p_a をキーワード a が示す記事として決定する．最後に，キーワード・記事ペア (a, p_a) それぞれについて，

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

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．

$$coherence(a, p_a) = \frac{1}{|S| - 1} \sum_{p_b \in S \setminus \{p_a\}} rel(p_b, p_a) \quad (3)$$

ピック抽出手法はどの言語に統一しても機能する．
*3 文献中では識別器を用いる TAGME も提案されている．

S は式 (2) の計算の対象となる全ての記事集合であり，式 (3) は候補の記事が互いに関連しているほど高い値となる．

オリジナルの TAGME は単一言語のテキストを対象とした手法であるため，本研究では，Wikipedia の言語間リンクによって TAGME を拡張する．これにより，任意の言語で記述されたテキストに英語の Wikipedia の記事を付与できる．具体的には，入力テキストに対してオリジナルの TAGME を適用し Wikipedia の記事を付与した後，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．また，エンティティリンキングの精度が検索精度に影響するという問題がある．試作システムでは，話題となっているトピックをより網羅的に取得するため，エンティティリンキングにおいて網羅性を重視する．すなわち，キーワードを表す正しい記事をより確実に付与できるように，キーワードに対して複数の記事をリンクする．これは，一方でノイズの増加を招く要因となりうるが，本研究が目的としているトピックの抽出においては，ノイズは統計的な情報によって抑えることが可能であると考えられる．提案システムでは，各キーワードについて式 (1) により得られるスコアの高い上位 ϵ% の記事のうち，確率 $Pr(p_a|a)$ の高い上位二つの記事をキーワード a が示す記事とする．最終的に，$\rho(a, p_a) > \rho_{NA}$ を満たす記事 p_a をキーワード a に付与するため，各キーワードに対して最大二つの記事が付与される．なお，パラメータ ϵ および ρ_{NA} の値として，文献 [4] を参考に，それぞれ 50% と 0.2 を用いた．

4.2 トピックグラフの構築

ソーシャルメディアにおいて，一つのテキストは基本的に一つのテーマ (単体あるいは関連する複数のトピックによって表現される) について言及されているため，同一のテキストに付与された記事に関する共起情報は，そのテーマにおける記事間の関連性の強さを表していると考えられる．そこで，記事をノード，同一のテキストに付与された記事の共起回数をエッジとしたトピックグラフ $G(V, E)$ を構築する．ここで，V はノード集合，E はエッジ集合である．これにより，任意の言語で記述されたテキストのトピック空間を英語の Wikipedia の空間に統一した状態で，トピック間の関連を言語別に表現できる．

4.2.1 記事間の共起情報の集計

記事の共起回数を集計する際，4.1 節で述べたように，テキスト中の一つのキーワードに対して複数の記事が付与される場合がある．エンティティリンキングでは，基本的にキーワードに対して正しいエンティティが一つだけリンクされるため，一つのキーワードに対して付与された二つのエンティティのうち，どちらか一方が正しいエンティティとする．すなわち，同一のキーワードに対して付与された記事同士を共起として扱わず，異なるキーワードに対して付与された記事間の共起についてのみカウントする．これにより，付与された記事のうち少なくとも一つが正しければ，そのテキストにおける記事間の共起情報が正しく得られるため，検索においてエンティティリンキングの精度の影響を抑えることができる．

4.2.2 静的な記事間関連度による共起情報の補足

ソーシャルメディアのテキストはテキスト長が短いため，同一のテキストに対して付与された記事の共起情報だけでは，トピック間のエッジがスパースになり，大部分のクエリに対して関連するトピック情報をほとんど提示できなくなる．この問題に対して，本研究では，記事間が持つ静的な関連度によって共起情報を補うことを考える．4.1 節で説明したように，試作システムでは，テキストに付与された記事が英語への言語間リンクを持たない場合についても，その記事は言語に特有なトピックとして扱うことを想定している．そこで，異なる言語の記事間の関連度計算に対応するため，Cross-Lingual Explicit Semantic Analysis (CL-ESA) [16] を採用する．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) [6] を Wikipedia の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，入力のテキスト中に出現する各語について，その語が出現する Wikipedia の記事を TF-IDF [15] や Okapi BM25 [14] などで重み付けしたベクトル (ESA ベクトル) を作成したあと，ベクトルの基底を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テキスト間の関連度計算を実現している．本研究では，CL-ESA ベクトルの作成に Okapi BM25 を用いた．

4.2.3 トピックグラフの作成

試作システムは，ソーシャルメディア上で話題となっているトピックの言語横断的な検索・可視化を目指しているため，どの記事を検索・可視化の対象とするかを決めるためのしきい値 τ を導入する．試作システムで対象とする全ての言語について，その記事が付与された回数の総和がしきい値 τ を超える場合，その記事 v を検索対象のトピック $v \in V$ として扱う．また，各言語 $l \in L$ におけるエッジ e_l の重みは次の式により計算する．記事 v_i, v_j 間のエッジ $e_l(v_i, v_j) = e_l(v_j, v_i)$ について，

$$e_l(v_i, v_j) = \log(Cooccur_l(v_i, v_j) + 1) + \log\left((1 + Sim(v_i, v_j)) \times \frac{\tau}{|L|}\right). \quad (4)$$

$Sim(v_i, v_j)$ はそれぞれの記事の CL-ESA ベクトルのコサイン類似度，$Cooccur_l(v_i, v_j)$ は言語 l における記事間の共起回数である．しきい値には，$\tau = 5000$ を用いた．

4.3 トピック情報の可視化

4.3.1 各言語におけるトピック情報の抽出

4.1 節および 4.2 節の処理によって得られた各言語のトピックグラフから，実際にどの記事が話題になっているのか，また，どの記事同士が関連して話題になっているのかを抽出するため，AffinityPropagation [5] を適用する．AffinityPropagation は，エッジの重みがノード間の関連度として定義されたグラフ (関連度行列) を入力としてクラスタリングを行い，クラスタおよび各クラスタを代表する exemplar を求めるクラスタリング手法である．k-means クラスタリングのような初期依存性がなく，preference と呼ばれるパラメータによりクラスタ数が自動的に決定される特徴を持つ．preference は，ノード自身へのエッジの重み (行列の対角成分) であり，値が大きいほどクラスタ数が増える傾向にある．文献 [5] では，preference の値として全エッジの重みの中央値が推奨されているが，各ノードで異なる preference を用いることで，ノードの exemplar としての選ばれやすさを個別に調整できる．また，responsibility と availability と呼ばれるノード間の関連性を表す値を用いることで，ノード自身の重要度やノード間の関係の強さを表すスコアを計算できる．

本研究では，エッジの重みとして共起回数を用いているため，ノード自身の重要度のスコアをその記事が実際にどれくらい言及されたを表す度合，ノード間の関係を表すスコアを記事同士が関連して言及された度合として考える．これらのスコアを用いることで，記事を指定したクエリに対して，関連して話題になっている記事のランキングをそれぞれの言語で作成できる．また，各記事の preference として記事の出現回数を用いることで，実際に多く言及された記事がランキングの上位に現れやすくなる．

AffinityPropagation では，responsibility と availability が収束するまで再帰的に計算する．ノード i，j について，responsibility $r(i,j)$ は，ノード j がノード i の exemplar としてどれほど適切であるかを表す値である．availability $a(i,j)$ は，ノード i がノード j を自身の exemplar として選択することがどれほど適切であるかを表す．responsibility と availability は初期値を 0 として，以下の式により値が収束するまで再帰的に計算する．

$$r(i,j) = (1-\lambda)\gamma(i,j) + \lambda r(i,j) \tag{5}$$

$$a(i,j) = (1-\lambda)\alpha(i,j) + \lambda a(i,j) \tag{6}$$

λ はダンピングファクタと呼ばれ，繰り返し計算における値の振動を抑制するパラメータである．$\gamma(i,j)$ および $\alpha(i,j)$ は，それぞれ繰り返し計算の各ステップにおける responsibility と availability の値であり，次式により計算する．

$$\gamma(i,j) = \begin{cases} e_l(i,j) - \max_{k\neq j}\{a(i,k) + e_l(i,k)\} & (i \neq j) \\ e_l(i,j) - \max_{k\neq j}\{e_l(i,k)\} & (i = j) \end{cases} \tag{7}$$

$$\alpha(i,j) = \begin{cases} \min\{0, r(j,j) + \sum_{k\neq i,j} \max\{0, r(k,j)\}\} & (i \neq j) \\ \sum_{k\neq i} \max\{0, r(k,j)\}\} & (i = j) \end{cases} \tag{8}$$

最終的に，ノード i の exemplar は収束値を用いて以下の式から計算する．

$$exemplar(i) = \arg\max_k \{r(i,k) + a(i,k)\} \tag{9}$$

式 (9) における $r(i,k) + a(i,k)$ は，記事 k の方が記事 i より話題であるという仮定の下で，記事同士が実際に関連して話題になっているかどうかを表すスコアであると言える．つまり，$r(i,k)+a(i,k)$ の値が大きければ，記事 k は記事 i より話題であり，かつ，記事 k と記事 i は関連して話題になっていることを意味する．ここで，ある記事 x について，$i = x$ として $r(i,k) + a(i,k)$ を全ての k について算出することで，各記事 k について記事 x より話題であり，かつ，記事 x とどの程度関連して話題になっているかを表すスコアを算出できる．同様に，$k = x$ として $r(i,k)+a(i,k)$ を全ての i について算出することで，各記事 i が記事 x とどの程度関連して話題となっているかを表すスコアを算出できる．試作システムにおいて記事 x がクエリとして与えられたとき，記事 x に関連して話題となっている記事が取得できればよいため，記事 x を除く全ての記事 y について

$$score_x(y) = \max\{r(x,y) + a(x,y), r(y,x) + a(y,x)\} \tag{10}$$

を計算することで，各言語において記事 x と関連して話題になっている記事 y のランキングを作成できる．

また，$r(i,k)+a(i,k)$ は記事 k について見たとき，記事 k が記事 i と比べてどれほど話題となっているかという度合を表している．つまり，ある記事 k に対して，$r(i,k)+a(i,k)$ について i の総和を取ることにより，記事 k が全ての記事の中でどれほど話題となっているかを表すランキングを求められる．このランキングは，試作システムにおいてクエリが空である場合の検索結果に等しい．

4.3.2 各言語のランキングの統合

4.3.1 項で求めた各言語におけるランキングを統合し，言語横断的に話題となっているトピックのランキングを作成する．試作システムでは，以下の二つの視点に基づく統合を行う．

4.3.2.1 全ての言語で共通して話題であるトピック

各言語のランキングにおいて，共通して上位に出現している記事を，全ての言語で共通して話題となっているトピックであると定義する．言語 $l \in L$ における記事 p のランクを $rank_l(p)$ としたとき，以下の式を用いて全ての言語で共通して話題となっている記事のランキングを作成する．

Vol.2014-DBS-160 No.11
Vol.2014-OS-131 No.1
Vol.2014-EMB-35 No.11
2014/11/18

$$score_{all}(p) = \frac{1}{|L|} \sum_{l \in L} \frac{1}{rank_l(p)} \qquad (11)$$

4.3.2.2 特定の言語でのみ話題であるトピック

各言語のランキングにおいて，ある言語を指定したときに，その言語のランキングにおいてのみ上位に出現する記事を，特定の言語でのみ話題となっているトピックであると定義する．以下の式を用いて言語 l でのみ話題となっている記事のランキングを作成する．

$$score_l(p) = \frac{1}{rank_l(p)} - \frac{1}{|L \setminus \{l\}|} \sum_{l' \in L \setminus \{l\}} \frac{1}{rank_{l'}(p)} \qquad (12)$$

4.3.3 トピックの可視化

4.3.2 項で求めた複数の言語間で統合されたランキングを，図 1 に示すように可視化することで，多言語なソーシャルメディアにおけるトピック情報を言語横断的に検索・比較可能にする．試作システムではまず，入力のクエリに対して式 (11) または式 (12) を用いてユーザが指定する形式のランキングを求め，ランキングの上位 k 個の記事を，クエリの記事と関連して話題となっている記事として選択する．この操作を，入力クエリから得られた記事に対して再帰的に適用することで，入力クエリに関するトピック情報を階層的に表現できる (図 1 では，$k = 5$ として 2 階層目まで表示している)．得られたトピック階層は，クエリの記事からのホップ数が少ない記事ほどクエリの記事と関連が強いトピックであることを意味している．

5. 予備実験

5.1 実験環境

試作システムの有効性を確認するために，4 ヶ国語 (英語，スペイン語，日本語，アラビア語) の Twitter のツイートを用いた予備実験を行った．本実験では，試作システムによってソーシャルメディア上のトピック情報を言語横断的に抽出・可視化できているかを，4.3 節で説明した二つの可視化手法それぞれについて検証する．具体的には，5 名の被験者に対して，試作システムを用いて英語の Wikipedia の記事をクエリとした検索を自由に行ってもらい，それぞれの可視化手法について以下の質問項目によるアンケート調査を行った．

(1) 全ての言語で共通して話題であるトピックの可視化について
  (a) クエリに関連したトピックが表示されているか．
  (選択肢：0=表示されていない，1=どちらかと言えば表示されていない，2=どちらかと言えば表示されている，3=表示されている)
  (b) 可視化前の各言語のランキングと比較して，全ての言語で共通して話題であるトピックが表示されているか．
  (選択肢：(1-a) と同じ)

表 1 実験に用いたデータセット

言語	ツイート数	キーワード数 /ツイート	付与記事数 /ツイート
英語	1,240,262	4.65	7.68
スペイン語	27,982	4.24	5.47
日本語	33,208	5.95	7.06
アラビア語	52,344	2.74	3.14

(2) 特定の言語でのみ話題であるトピックの可視化について
  (a) クエリに関連したトピックが表示されているか．
  (選択肢：(1-a) と同じ)
  (b) (2-a) で 0 または 1 と回答した場合：その理由は，可視化前の各言語のランキングと比較して，(i) クエリに対して言語特有な話題が正しく取得できていない，(ii) クエリに対して言語特有な話題がない，のどちらか．
  (選択肢：0=(i)，1=どちらかと言えば (i)，2=どちらかと言えば (ii)，3=(ii))
  (c) (2-a) で 2 または 3 と回答した場合：可視化前の各言語のランキングと比較して，その言語のみで話題となっているトピックが表示されているか．
  (選択肢：(1-a) と同じ)

質問項目 (1) では，クエリを自由に指定して検索してもらい，各クエリに対する結果について，(1-a)，(1-b) の回答を集めた．質問項目 (2) では，英語，スペイン語，日本語，アラビア語のそれぞれの言語について，言語を指定した状態で自由に検索してもらい，各クエリに対する結果について，(2-a)，(2-b) または (2-c) の回答を集めた．

本実験では，試作システムによってトピック情報を抽出および検索・可視化できているかを簡便に確認するために，「WorldCup」という一つのテーマに関するツイートをデータセットに用いた．具体的には，2014 年 7 月 2 日から 7 月 14 日にかけて収集したハッシュタグ「#WorldCup」を含む英語，スペイン語，アラビア語，日本語の 4 ヶ国語のツイートを用いた．データセットの作成手順として，1)Streaming API を用いてハッシュタグ「#WorldCup」を含むツイートを収集，2) リツイート，URL の除去，3) ハッシュタグの「#」のみを削除，の各処理を行ったあと，全てのツイートについて図 2 に示した流れに従ってトピック抽出を行った．データセットの統計情報を表 1 に示す．

5.2 実験結果

アンケート結果を表 2 に示す．表 2 では，列ごとに被験者のアンケート結果を表しており，各セル内の値は質問項目に対する各被験者の回答の平均値を表している．

はじめに，質問項目 (1)「全ての言語で共通して話題であるトピックの可視化手法」の結果について考察する．質問項目 (1-a) について，全ての被験者が検索したクエリに

ついて関連したトピックが表示されていたと回答している．これは，試作システムにおいて，同一のテキストに対してリンクされた記事同士の共起情報を用いてトピックグラフを構築することにより，関連して話題となっている記事間の関係をうまく抽出できていることを示している．質問項目 (1-b) について，平均値の最小値が被験者 A の 2.4 であることから，実際に被験者が各言語におけるトピックのランキングを比較した場合に言語間で共通して話題であると考えるトピックと，試作システムによって可視化したトピックが類似していることがわかった．質問項目 (1) の結果から，試作システムは，クエリに関連したトピックを検索でき，かつ，全ての言語で共通して話題であるトピック情報を可視化できていると考えられる．

次に，質問項目 (2)「特定の言語でのみ話題であるトピックの可視化手法」の結果について考察する．質問項目 (2-a) について，ほとんどの項目で平均値が 2 以上となっており，言語に特有なトピックを可視化する場合においても，クエリの記事に対して関連して話題となっている記事を検索できていると言える．言語に特有なトピックを検索する場合，クエリによってはそもそも言語に特有なトピックが無い場合がある．試作システムでは，4.3 節で説明した手順に従って作成したランキングで上位となっている記事を，指定した個数だけ表示する形式をとっている．このような場合，試作システムは，言語に特有でなく，また，話題になっていないトピックを表示することになる．実際に，被験者 A，D は一部のクエリについて，関連したトピックが表示されておらず，また，そのクエリについて対象の言語のみで話題となっているトピックが無かったことを回答している (被験者 A の (2-b)AR および被験者 D の (2-b)EN の回答)．特定の言語でのみ話題であるトピックを可視化する場合，実際に求めたランキングの中に可視化手法の目的に合った記事が含まれているかどうかを判定し可視化する必要がある．

一方，言語に特有なトピックが抽出できていないという回答もあった (被験者 A の (2-b)EN および JA の回答)．これらの回答で使用されたクエリについて，実際に可視化されたトピック情報を確認したところ，言語に特有なトピックが表示されているが，話題となっていないトピックも表示されていた．また，各言語のランキングにおいて，クエリと関係がないと思われる記事が上位に多数含まれている場合もあった．これは，エンティティリンキングによりキーワードに対して誤った記事が付与されたことにより，誤った共起が多く発生したためであると考えられる．今後は，エンティティリンキングの精度についても改善していく必要がある．

6. まとめと今後の課題

本研究では，多言語なソーシャルメディアにおけるトピック情報を言語横断的に抽出・可視化するシステムを試作した．試作システムでは，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対してエンティティリンキングを行い，テキストに出現するキーワードを英語の Wikipedia の記事にリンクした後，英語の Wikipedia の記事をノード，記事同士の共起関係をエッジとしたグラフとして表現する．これにより，トピックの言語空間を統一した状態で，トピック間の関連を言語別に表現できる．また，Wikipedia の記事をクエリとして与えることで，その記事に関連して話題となっているトピックを言語横断的に可視化できる．Twitter のデータを用いた予備実験により，試作システムが複数の言語で共通したトピックや特定の言語でのみ話題となっているトピックを可視化できていることを確認した．

表 2 各被験者の回答

質問項目/被験者	A	B	C	D	E
(1-a)	3	3	3	3	3
(1-b)	2.4	2.8	2.7	3	3
(2-a) EN	1.6	3	2.6	1.6	2.5
(2-a) ES	2.6	3	3	3	3
(2-a) JA	2.3	3	3	3	2.5
(2-a) AR	2.2	3	3	3	2.5
(2-b) EN	0.7	-	-	2.5	-
(2-b) ES	-	-	-	-	-
(2-b) JA	1	-	-	-	-
(2-b) AR	2	-	-	-	-
(2-c) EN	2.5	2	2.4	2.7	2.5
(2-c) ES	2.8	1.8	2.6	2.8	2
(2-c) JA	2	2	2.2	2.2	2
(2-c) AR	3	2.2	2.8	2.8	2.5

今後の課題は，実験結果から得られた知見をもとにトピック抽出手法や可視化手法の改善点を洗い出し，評価方法のデザインを検討することである．トピック抽出手法における改善点としては，エンティティリンキングの精度向上が挙げられる．今回の試作システムでは，エンティティリンキングの候補となる記事に制限を設けていなかったため，エンティティの粒度の不一致が発生していた．例えば，「FIFA World Cup」の記事が付与できれば良いキーワードに対して，過去に行われた特定の年度のワールドカップに関する記事が付与される問題が生じていた．このような問題に対しては，Wikipedia のカテゴリ情報等からエンティティリンキングの候補となるエンティティをある程度集約することが有効であると考えられる．また，今回の予備実験ではワールドカップというテーマを絞ったツイートを用いたが，今後は，トピックを制限せず Twitter から得られる任意のトピックに関するツイート集合に対してトピック抽出手法を適用し，システムの有効性を確認する予定である．

可視化手法についての改善点としては，現在はトピック

間の単純なつながりだけを表現することで言語横断的な検索・可視化を実現しているが，今後は，詳細なトピック情報を同時に提示することで，より効果的な可視化を目指す．例えば，各トピックの話題の度合やトピック間の関係の強さをノードの大きさやエッジの太さなどで視覚的に表現したり，そのトピックについて言及しているソーシャルメディアのテキストを併せて提示したりすることを考えている．また，言語横断的にトピックを検索・可視化する際に，どのような要求があるのか，またその要求を満たす可視化手法について詳細に検討する．

トピック抽出手法と可視化手法の改善と同時に，トピック抽出手法の精度面の評価と可視化手法に関する評価のデザインについて検討する．トピック抽出手法の精度に関する評価では，抽出したトピック情報とそれらのトピックについて言及しているソーシャルメディアのテキストを比較し，テキストの内容に合ったトピック情報が得られているかを検証することを考えている．実際に，既存研究 [8], [20] では，提案手法によって抽出したトピック情報について，被験者が手動で抽出したトピック情報とどれだけ一致しているかをみることで精度に関する評価を行っている．可視化に関する評価においては，様々な可視化の要求について，提案した可視化手法によりその要求を満たすことができているかを検証する．

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．

参考文献

[1] Blei, D. M., Griffiths, T. L. and Jordan, M. I.: The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies, *Journal of the ACM*, Vol. 57, No. 2, pp. 7:1–7:30 (2010).

[2] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[3] Boyd-Graber, J. and Blei, D. M.: Multilingual Topic Models for Unaligned Text, *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 75–82 (2009).

[4] Ferragina, P. and Scaiella, U.: Fast and Accurate Annotation of Short Texts with Wikipedia Pages, *IEEE Software*, Vol. 29, No. 1, pp. 70–75 (2011).

[5] Frey, B. J. and Dueck, D.: Clustering by Passing Messages Between Data Points, *Science*, Vol. 315, pp. 972–976 (2007).

[6] Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 1606–1611 (2007).

[7] Jagarlamudi, J. and Daumé, H.: Extracting Multilingual Topics from Unaligned Comparable Corpora, *Proceedings of European Conference on Advances in Information Retrieval (ECIR)*, pp. 444–456 (2010).

[8] Kang, W., Tung, A. K., Zhao, F. and Li, X.: Interactive Hierarchical Tag Clouds for Summarizing Spatiotemporal Social Contents, *Proceedings of International Conference on Data Engineering (ICDE)*, pp. 868–879 (2014).

[9] Li, W. and McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations, *Proceedings of International Conference on Machine Learning (ICML)*, pp. 577–584 (2006).

[10] McNamee, P., Mayfield, J., Lawrie, D., Oard, D. W. and Doermann, D. S.: Cross-Language Entity Linking, *Proceedings of International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 255–263 (2011).

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

[12] Milne, D. and Witten, I. H.: An Effective, Low-cost Measure of Semantic Relatedness Obtained from Wikipedia Links, *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence*, pp. 25–30 (2008).

[13] Ni, X., Sun, J.-T., Hu, J. and Chen, Z.: Cross Lingual Text Classification by Mining Multilingual Topics from Wikipedia, *Proceedings of ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 375–384 (2011).

[14] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M. et al.: Okapi at TREC-3, *NIST SPECIAL PUBLICATION SP*, pp. 109–109 (1995).

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

[16] Sorg, P. and Cimiano, P.: Cross-lingual Information Retrieval with Explicit Semantic Analysis, *Working Notes for the CLEF 2008 Workshop* (2008).

[17] Wang, Y.-C., Wu, C.-K. and Tsai, T.-H. R.: Cross-language and Cross-encyclopedia Article Linking Using Mixed-language Topic Model and Hypernym Translation, *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 2, pp. 586–591 (2014).

[18] Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E.-P. and Li, X.: Topical Keyphrase Extraction from Twitter, *Proceedings of Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, Vol. 1, pp. 379–388 (2011).

[19] Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H. and Li, X.: Comparing Twitter and Traditional Media Using Topic Models, *Proceedings of European Conference on Advances in Information Retrieval (ECIR)*, pp. 338–349 (2011).

[20] Zhu, X., Ming, Z.-Y., Zhu, X. and Chua, T.-S.: Topic Hierarchy Construction for the Organization of Multi-source User Generated Contents, *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrievalonference on Research and Development in Information Retrieval*, pp. 233–242 (2013).

[21] 寛之中崎，真理子川場，大輔横本，武仁宇津呂，知宏福原：多言語 Wikipedia エントリを知識源とする特定トピックの日英ブログサイト検索と日英対照ブログ分析，人工知能学会論文誌，Vol. 25, No. 5, pp. 613–622 (2010).