

次投稿予測モデルを用いた想起関係の自動抽出

内田 脩斗^{1,a)} 吉川 大弘^{1,b)} 古橋 武^{1,c)}

受付日 2019年8月22日, 再受付日 2019年10月16日,
採録日 2020年1月22日

概要: 想起とは, ある概念が別の概念をどの程度思い浮かばせるかを表した重み付き情報であり, 様々なアプリケーションへの応用が期待されている. 一般に, 人間の知覚や経験に強く依存する意味関係は, 構造的な意味関係と比べ, その抽出が困難であると考えられており, 想起関係においても研究の余地を多分に残している. そこで本研究では, 教師なし学習による想起関係の自動抽出を試みる. 従来のアプローチでは, 単語の共起情報を利用した手法が報告されているが, 想起関係は方向性のあるデータであるため, 共起情報のみでは不十分と考えられる. 本論文では, コーパス内投稿データの話題遷移情報を利用した想起関係の自動抽出手法を提案する. また, 英語と日本語による想起評価データを用いた実験を行い, 従来手法との性能比較結果を示す. 実験の結果, 提案モデル単体では従来モデルと同程度以上の性能であること, また, 両者をアンサンブルすることで性能が大幅に向上することを報告する. これより, 従来では用いられなかった想起の方向性を考慮した想起抽出モデルの作成が重要であることを示唆する.

キーワード: 想起, 教師なし学習, トピックモデル, 状態遷移, ニューラルネットワーク

Automatic Extraction of Evocations by Predicting Model of Next Posts

SHUTO UCHIDA^{1,a)} TOMOHIRO YOSHIKAWA^{1,b)} TAKESHI FURUHASHI^{1,c)}

Received: August 22, 2019, Revised: October 16, 2019,
Accepted: January 22, 2020

Abstract: Evocation is weighted information that expresses how much a concept evokes another, and its application to various tasks is expected. In general, the extraction of semantic relationships strongly depending on human perception and experience is considered to be more difficult than the extraction of structural semantic relationships, and there is much room for researches in evocation relationships as well. So, in this paper, we try an extraction method of evocation relations using unsupervised learning. Conventionally, a method using co-occurrence relation of words has been reported. However, since evocation relations are directional, the co-occurrence information alone is considered to be insufficient. Therefore, this paper proposes an extraction method of evocation relation using the transition information of post data in corpus. We also conduct the experiments using evocation evaluation data in English and Japanese, and we compare the performance of proposed method with that of the conventional method. As a result, it shows that the performance of the proposed model is as high as or higher than that of the conventional one, and it can be further improved by their ensemble. This result suggests that it is important to create an evocation extraction model that takes into account the direction of evocation, which was not considered in the conventional method.

Keywords: evocation, unsupervised learning, topic model, state transition, neural network

1. はじめに

膨大なテキストデータが日々蓄積されている昨今, ビッグデータから有用な知見を得るための情報抽出技術に関する研究がさかんに行われている [1], [2]. 特に近年では, 単語の意味的な関係性を多次元ベクトルとして学習する

¹ 名古屋大学大学院工学研究科
Graduate School of Engineering Nagoya University, Nagoya,
Aichi 464-8603, Japan

a) uchida@cplx.cse.nagoya-u.ac.jp

b) yoshikawa@cse.nagoya-u.ac.jp

c) furuhashi@cse.nagoya-u.ac.jp

言語モデルが数多く報告されており [3], [4], [5], [6], 様々な NLP タスクへの応用が進んでいる。また、単語どうしの意味関係に限らず、視覚に代表される人間の知覚情報を取り込んだ意味表現の獲得に関する研究も報告されており [7], [8], さらに発展が期待されている。特に本研究では、人間の知覚情報の一種である“想起”に焦点を当てる。

想起とは、“ある概念が別の概念をどの程度思い浮かべさせるか”を表した重み付き情報として定義されており [9], 以前より心理学の分野で広く用いられている [10], [11], [12]. また、この想起関係情報は、検索アルゴリズムや推薦システムなどのパーソナライズ技術や宣伝広告などのマーケティング戦略への適用など、様々なアプリケーションへの応用が期待されている [13]. 一方で、人間の知覚や経験に強く依存する意味関係は、構造的な意味関係（上位・下位概念、同義語など）と比べ、その抽出が困難であると考えられており [14], [15], 研究の余地を多分に残している領域である。

本論文では、ソーシャルメディアを用いた想起関係の自動抽出に注目し、教師なし学習の適用可能性を報告する。従来のアプローチでは、LSA (Latent Semantic Analysis) [16] を用いて単語ベクトルをテキストコーパスより学習し、単語間の cosine 類似度を想起度合いとして用いる手法が最も効果的であると報告されている [9]. これは、単語の共起情報を用いた想起関係抽出手法であるが、一方で想起関係は方向性のあるデータである。たとえば、[17] の調査では、*beer* から想起されやすいものは *glass* であるが、逆方向である *glass* から *beer* は相対的に想起されにくいと報告されている。よって、単語の共起情報のみを用いる手法では、単語間の想起関係を適切にとらえるためには不十分であると考えられる。

そこで著者らは、コーパス内投稿データの話題遷移情報を利用した想起関係の自動抽出手法を提案する。ここでは、同一ユーザの投稿は過去の投稿に依存しやすい、すなわち、過去の投稿を手がかりとして次の投稿を行いやすいという投稿間想起仮説をおく。具体的には、投稿データの遷移関係を学習する次投稿予測モデルを構築することで話題遷移関係を抽出し、その関係性を想起関係と見なす手法である。これにより、想起の方向性を考慮した想起関係抽出モデルとなることが期待される。

特に本論文の貢献は、より精度の高い想起抽出モデルの考案、また、投稿間想起仮説の有用性の示唆である。実験では、文献 [9] の想起データ（英語）と独自に収集した想起データ（日本語）を用いた想起度合いの順位評価を行い、その結果を 4.5 節に示す。この結果より、後者の想起データにおいて、提案モデルの予測精度が従来モデルの予測精度と同程度以上であることを報告する。さらに、改良モデルとして両手法による予測結果をアンサンブルした場合に予測精度が大幅に向上することを示す。これは、共起情報

に強い従来モデルと、方向情報に強い提案モデルが補完関係にあることに起因すると考えられる。つまり、従来では用いられなかった概念の方向性を想起関係抽出モデルに取り込むことの重要性を示唆しており、より頑強な想起抽出モデルを構築していくことの必要性を表している。

2. 関連研究

想起に関する研究は古くから行われており、特に心理学の分野において、人の想起プロセスに関する研究が報告されている [10], [11], [12], [18]. また、Boyd-Graber ら [9] は、概念単語間の想起度合いのアノテーションを行い、初めて大規模な想起データを作成した。この想起データは、WordNet [19] における 1,000 のコア概念単語を対象とし、ランダムに抽出された約 12 万件の概念ペア間の想起の程度を、少なくとも 3 名以上の被験者に評定させたものである。想起度合いは、0 から 100 の整数でスコアリングされており、何らかの想起関係があることを意味する 0 より大きいと評価されたペア数は、全体の約 33% と非常に偏りのあるデータとなっている。さらに、想起関係の自動抽出手法についても言及しており、WordNet ベースの手法 [20], [21] や LSA ベースの手法 [16] の適用を試みている。そこでは、スピアマンの順位相関係数 ρ による評価の結果、両手法とも同程度の精度 ($\rho = 0.131$) となったことを報告している。また、Nikolova ら [22] は、Boyd-Graber らの想起データ収集手法では、想起スコアの低いデータが多く存在し、スコアの高いデータの収集が非効率的であることに對して、上記のデータを用いて AdaBoost により単語ペアの想起スコアを予測し、スコアの高い想起ペアを採用してアノテーションを行うことで、全体の約 60% に想起関係のある想起データを作成したことを報告している。さらに、Ma ら [15] は、Free Word Association data [23] を用いた想起データ収集手法を提案し、Nikolova らの手法との比較を行っている。一方で林 [13] は、想起度合いの予測に対して回帰予測モデルの適用を試みている。この実験では、様々な単語の特徴量を考案し、ニューラルネットワークによる想起度合いの回帰予測を行っており、その結果、特徴量として Word2Vec [3] による単語ベクトル間 cosine 類似度を用いた場合に $\rho = 0.184$, 全 8 種の特徴量を組み合わせた場合に $\rho = 0.400$ となったことを報告している。

想起関係は、人の経験や生活環境に強く依存する関係性である [13], [15] ととらえられているため、より個々の属性を考慮した大量のデータ収集が必要であり、その収集には多大な労力やコストを必要とする。よって、自動で生成・抽出が可能な手法が求められている。一方で、教師あり機械学習モデルでは、アノテーション付きの膨大なデータセットが必要であるため、コーパスなどの大規模データから自動的に想起関係が抽出できるモデルを構築することが重要である。そこで本論文では、教師なし機械学習による

表 1 投稿間想起関係の評価例

Table 1 Examples of relationship between post interval and evocation.

想起関係あり とした人数	過去の投稿	未来の投稿
3	乳児湿疹の子用掻き難い服作って	携帯出来る使い捨て哺乳瓶ってないのかな
2	レジの処理が遅い方がいます。 隣のレジの方が早い。急いでるわけではないが、不満。	〇〇駐車場に別なスーパーを利用している人が停めている
1	文が英語だけでなく日本語もしっかりやってほしい	充電が7%あるのに充電がおちて そのままつけたら充電が20%残ってる
0	時間外未払いが日常茶飯事におこなわれている	路側の草がたくさん生えているので通行の妨げになる

想起抽出手法について述べる。

3. 提案モデル

本章では、投稿間想起仮説に基づいた次投稿予測モデルによる想起関係の抽出手法について論じる。

3.1 投稿間想起仮説

本論文では、不満買取センターと呼ばれる Web サービスにて収集された不満調査データセット [24] を用いる。これは、ユーザの感じた不満を自由記述形式で収集するサービスで、豊富な投稿データやメタデータ (居住地, 性別など) を用いることが可能なデータセットである。また、ユーザは投稿に対して金銭的報酬が得られるため、データの品質は一般的なソーシャルメディアと比べ高いと思われる。

本論文では特に、本データにおける各ユーザの投稿の系列情報に注目する。不満調査データセットの場合、ユーザは不満を投稿するという状況であるため、自身の体験をもとに投稿を発信する。また、投稿に対して報酬が与えられるため (一般的に投稿内容が詳細であるほど高額となる)、不満を体験した時点で投稿を発信するというより、過去の体験を思い出して投稿を発信すると考えられる。実際に前後関係にある投稿データの投稿間隔時間を確認したところ、全体の約 50% が 20 分以内に次の投稿が行われていた。つまり、リアルタイムに情報が発信されるというよりは、記憶を掘り起こして投稿を発信するユーザが多いコーパスであると考えられる。また、記憶を掘り起こすのに効果的とされている手法の 1 つに手がかり再生がある。これは、何らかの手がかりを利用して思い出す行為のことであり、すなわち、想起であるといえる。つまり、本データではきわめて短期間に次の不満投稿が行われているため、自身が過去に投稿した内容を考慮して、次の投稿を発信する可能性が高いと思われる。これにより、投稿間隔が短いほどそれらの関係性が強いと考えられ、また、投稿間隔が空いたとしても過去の投稿を手がかりにすることもありうる。

これらをふまえて、提案モデルでは、「あるユーザの投稿は自身の過去の投稿に依存しやすい」という投稿間想起仮説をおく。これは、過去の投稿を手がかりとして次の投稿

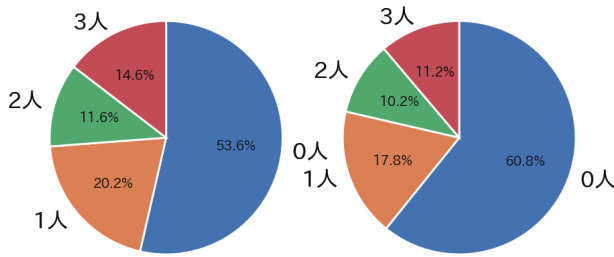
を発信しやすい (想起しやすい) と考えることができる。なお簡単のため、本研究ではマルコフ性 (次の投稿は過去の最新の投稿のみに依存する) を仮定している。これらを仮定すると、投稿内容 (話題) の遷移関係を、話題 (概念単語) の想起関係と見なすことが可能となり、想起関係の自動抽出に役立つ可能性がある。

そこでまず、投稿間想起仮説の検証を目的に、被験者 (男子大学院生 3 名) に対して、実際のデータ内の投稿間関係と想起関係の関係性を確認する実験を行った。実験内容は、ある投稿が以前の投稿から連想可能か否かを判定するものである。対象とするユーザは、投稿数が 2 回のみユーザとした。また、投稿間隔と想起関係には依存関係が考えられるため、データは投稿間隔が 20 分以内/1 日以上投稿ペアと、ランダムに抽出した投稿ペアをそれぞれ 500 セット抽出し、計 1,500 セットの評価を行った。

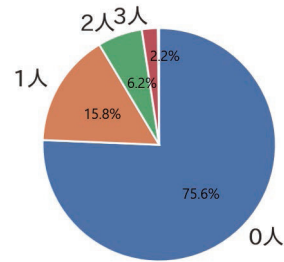
表 1 に、本実験により評定された実際の例を示し、図 1 に、投稿間に想起関係があるとした被験者数の割合を示す。図 1 より、投稿間隔が短い方が、連想関係があると判定されたデータが多く存在することが確認できる。また、図 1 (c) と比較すると、前後関係にある投稿間には想起関係が存在しやすいことが確認できる。これにより、前述のとおり、「あるユーザの投稿は自身の過去の投稿に依存しやすい」といえるだろう。ただし本実験では、被験者は限りある投稿データの情報から、想起関係があるか否かをとらえており、誤りや見逃しが多分に含まれている可能性があると考えられ、他のデータでも検証する必要があると思われる。

3.2 次投稿予測モデル

想起関係を抽出するために、まず出現する単語を概念系に落とし込むため、概念抽出手法として LDA (Latent Dirichlet Allocations) [25] を用いて、トピック分布の学習を行う。これにより、式 (1) を用いることで投稿文書のトピック分布を推論し、文書ベクトルとして利用することができる。ただし、 t は投稿文書、 w は投稿内出現単語、 z はトピック分布を表している。



(a) 投稿間隔：20 分以内 (b) 投稿間隔：1 日以上



(c) ランダムな投稿ペア

図 1 投稿間隔と想起関係の関係 (ラベル：想起関係ありとした人数)
Fig. 1 Relationship between post interval and evocation (Label: The number of subjects who evaluated that they have an evocation relationship).

$$\begin{aligned}
 p(z|t) &= \sum_w p(z|w)p(w|t) \\
 &= \frac{p(w|z)p(z)}{\sum_{z'} p(w|z')p(z')} p(w|t)
 \end{aligned} \tag{1}$$

さらに、投稿の前後関係を表した投稿ベクトルのペアデータを作成し、次投稿の投稿ベクトルを予測するモデルを学習する。これにより、話題間遷移を示す状態遷移行列を生成することが可能となる。よって、投稿間想起仮説に基づき、状態遷移行列を概念間想起関係と見なすことで、想起関係の抽出が達成される。また、具体的な予測モデルについては、TM-LDA と MCNN-LDA の 2 種の利用を検討する。

3.2.1 Temporal - LDA (TM-LDA)

次投稿予測モデルとして提案されている手法の 1 つに、TM-LDA [26] がある。TM-LDA では、LDA により生成されるトピック分布を用い、マルコフ性を仮定することで、文書の前後関係を表現する状態遷移行列 T を式 (2) により獲得する。

$$T = (A^T A)^{-1} A^T B \tag{2}$$

このとき、 A は過去の投稿データをトピックベクトル化した行列、 B は未来の投稿データをトピックベクトル化した行列を表している (図 2)。ただし、 A, B の各行ベクトルは式 (3) を満たす必要がある。ここで、 w_d はあるトピックベクトルの要素を表している。

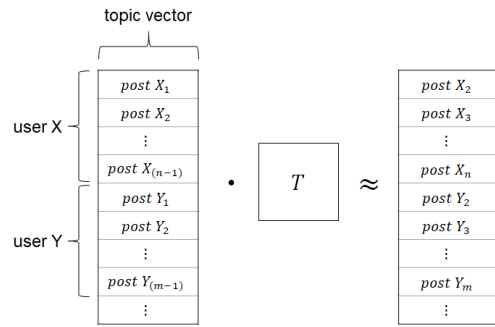


図 2 TM-LDA の構成

Fig. 2 Configuration of TM-LDA.

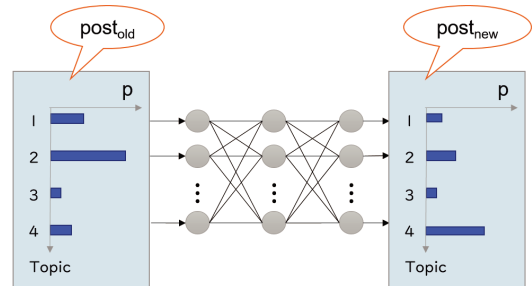


図 3 MCNN-LDA の構成

Fig. 3 Configuration of MCNN-LDA.

$$\sum_{d=1}^K w_d = 1 \quad (0 \leq w_d \leq 1) \tag{3}$$

これにより、学習される T はトピックからトピックへの遷移確率を表すことができ、その確率をトピック概念間の想起度合いと見なすこととする。

3.2.2 Markov Chain Neural Network - LDA (MCNN-LDA)

本手法では、ニューラルネットワーク (以下、NN) による次投稿予測モデルを構築する。つまり、図 2 の状態遷移行列 T に対応する部分に、NN を適用する (図 3)。これにより、NN による状態遷移行列の表現が可能のため、Markov Chain Neural Network -LDA (以下、MCNN-LDA) と呼称する。MCNN-LDA は、TM-LDA と比べてモデルの柔軟性の高さが一番のメリットである。まず、TM-LDA では T の表現能力が (トピック数, トピック数) の行列と限定されるため、これを拡張することによる性能向上が期待できる。また、TM-LDA では入力 A と出力 B の行列サイズが一致することが前提にあるため、入力データの柔軟性に欠けるという点がある。そのため、投稿データに出身地や年齢などのメタデータが存在する際に、容易にモデルに組み込むことができない。2 章でも述べているとおり、想起関係は個人の経験・生活環境に強く依存すると考えられるため、属性ごとに柔軟に想起関係を再定義できることはモデルの重要な要素となりうる。MCNN-LDA では入力データの型制約がなく、メタデータに対応した次元を新た

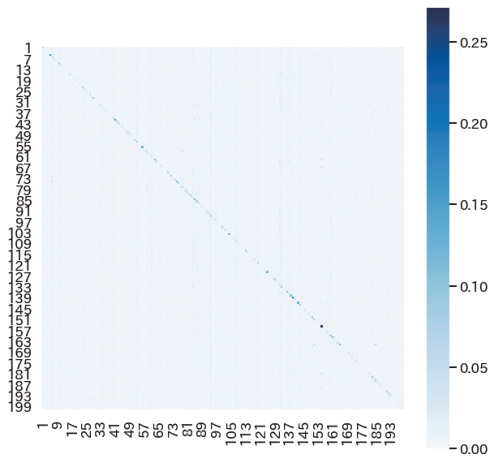


図 4 状態遷移行列の可視化
Fig. 4 Visualization of state transition matrix.

に入力層に付与することで、個々の想起関係を獲得することが可能となると考えられる。また MCNN-LDA は、図 3 のように出力ベクトルがトピックベクトル（確率分布）という制約を保持している。つまり、式 (3) を満たす必要がある。そのため、出力層には softmax 関数（式 (4)）を導入することで、予測値を確率分布として扱えるようする。

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{d=1}^K \exp(x_d)} \quad (4)$$

ただし MCNN-LDA では、予測モデルがブラックボックスであるため、直接的に状態遷移行列 T を観測することはできないという問題がある。そこで、状態遷移行列を擬似生成することを考え、NN の入力値として単位行列 I を導入する。このとき、ある行ベクトルに注目すると、これはあるトピックからそれぞれのトピックへの遷移確率を予測する演算となる。つまり、それらの予測結果の行列は状態遷移行列 T として扱うことが可能となる。また、TM-LDA と同様に、トピック間遷移確率を概念間の想起関係と見なすこととする。

図 4 は、実際に上記方法により生成したトピック間状態遷移行列である。使用した投稿データは、表 2 と同様である。図 4 では、縦軸が過去のトピック、横軸が未来のトピックに対応している。これより、行列の対角成分に高い確率が与えられていることが確認できる。これは、同一トピックへの遷移確率が高いことを表しており、投稿の前後で類似の内容を投稿する傾向が強いといえる。なお TM-LDA においても、同様の行列が生成される。

4. 想起評価実験

本章では、各モデルにより実際に抽出された想起関係と評価データとの比較実験を行い、その結果について考察する。

表 2 不満調査データセットのデータ詳細

Table 2 Data details of FKC corpus.

対象期間	2015/03/18-2017/03/12
ユーザ数	64,196
学習用投稿ペアデータ数	2,965,783
ボキャブラリ数	96,720

表 3 MCNN-LDA のパラメータ設定

Table 3 The parameters of MCNN-LDA.

パラメータ	値
層数	3
次元数	すべて 200
Optimizer	Adam
初期学習率	0.001
活性化関数	relu
損失関数	MSE
epochs	15
ミニバッチサイズ	128
終了条件	early stopping

4.1 想起関係抽出用データ

表 2 に、想起関係抽出に用いた不満調査データセットのデータ詳細を示す。本実験では、投稿数が 3~1,000 のユーザを対象とした。また、居住地・職業・性別の属性に欠損のないユーザを抽出した。さらに前処理として、投稿データには、ストップワードの除去と名詞の抽出、低頻度語の除去を行った*1。さらに前処理後のデータを用いて、3.2 節に示した想起関係抽出モデルの学習を行った。

4.2 MCNN-LDA パラメータ

本実験で設計した NN のパラメータを表 3 に示す。本実験では、3 層のネットワークを用い、入力・出力のトピック数、および中間層の次元数は 200 とした。

4.3 評価用想起データ

4.3.1 PWN-Evocation

概念単語間の想起関係を評価付けしたデータセットとして、文献 [9] の PWN-Evocation（以下 PWN-E）を用いた*2,*3。これは、2 章でも紹介した WordNet を用いた英語の想起データセットである。その内容としては、ある単語（e.g., job）からある単語（e.g., potential）を想起する度合いを 0 から 100 の整数でスコアリングが行われているものである。また、その意味合いとして、0：まったく想起されない、25：離れた想起関係がある、50：中程度の想起関係がある、75：強い想起関係がある、100：即座に想起される、と表現できる。また、全データのスコア分布を

*1 実装には、Python3.6.7, mecab0.996.1, gensim3.6.0, Keras2.2.4 を用いた。

*2 <http://wordnet.cs.princeton.edu/downloads.html>

*3 文献 [22] の想起データでも同様の実験を行ったが、類似した結果となったので割愛する。

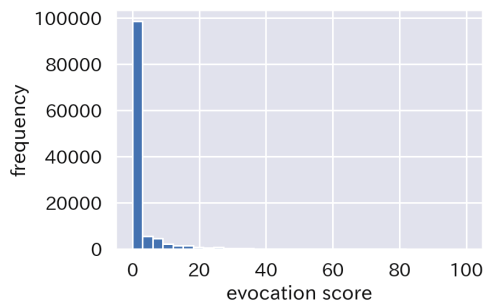


図 5 PWN-Evocation におけるスコア分布
Fig. 5 Score distribution in PWN-Evocation.

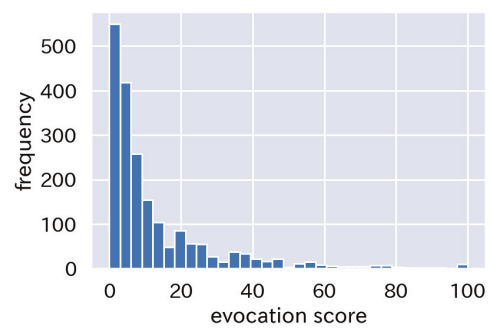


図 6 Japanese-Evocation におけるスコア分布
Fig. 6 Score distribution in Japanese-Evocation.

図 5 に示す。

ただし本論文では、日本語の不満調査データセットを利用しているため、評価の際に単語の翻訳が必要である。そのため、PWN-E 内の単語は Weblio 翻訳^{*4}を用いて日本語に変換したのち、目視確認による修正を行うことで、日本語単語との紐付けを行った。仮に、複数の英単語に対して同一の日本語単語が紐付けられた場合は、それらすべてのペアデータを許容した。また、LDA では各トピックに全語彙と単語出現確率が与えられるため、単語は単語出現確率最上位のトピックに所属するものとした。

4.3.2 日本語想起データの収集

4.3.1 項では、既存の想起データセットを紹介したが、これは英語の想起データであり、日本語コーパスから抽出された想起関係とはマッチしない可能性がある。そこで、被験者（男子大学院生 3 名）による日本語の想起データ（Japanese-Evocation, 以下 J-E）の作成を行った。想起データ収集では Boyd-Graber らの方法を参考にし、以下の指示を行った。

- (1) 想起とは、“ある概念が別の概念をどの程度思い浮かばせるか”を表しています。
- (2) base 単語を元に、next 単語をどの程度想起させるかを 0 から 100 の整数で評定してください。
- (3) 想起度合いの目安：0：まったく想起されない、25：離れた想起関係がある、50：中程度の想起関係がある、75：強い想起関係がある、100：即座に想起される
- (4) 想起関係は必ずしも対称とは限りません。

また、想起ペアデータの単語となる概念単語は、評価が行いやすいように不満調査データセットをベースに選定し、190 単語の中からランダムに想起ペアを 2,000 件作成し、アノテーションを行った。作成した J-E の想起スコア分布を図 6 に示す。

4.3.3 想起データの比較・考察

図 5、図 6 より、PWN-E と J-E では、スコア分布に大きな差があることが確認できる。この原因として、まず、PWN-E では評価数が約 12 万件と多く、評価者の集中力が続かず、単調なスコア付けをする傾向があったと考えら

れる。また、PWN-E では具体的な単語 (e.g., cane, cash register) が多く、J-E では抽象的な単語 (e.g., ペット, 仕事) が多いことがあげられる。これは、J-E では、抽象的な単語を選択して想起データ作成に利用しているためである。このため、J-E の想起ペアの方がイメージが広がりやすく、想起関係が生じやすいことで、スコア分布に広がりやすかったものと考えられる。

4.4 評価フロー

提案手法における評価フローを以下に述べる。

- (1) あらかじめ、トピック内上位何単語 (Top-n) を対象とするか設定する。
- (2) 評価可能な想起ペアデータを抽出する。
- (3) 単語間の想起度合いを予測する。
- (4) 評価用想起データのスコアと比較し評価値を算出する。

本実験では評価指標として、文献 [9], [13] でも用いられているスピアマンの順位相関係数を用いた。これは、想起データのスコア分布に偏りがあるためである。想起度合いの算出は、学習済みのトピック遷移確率 $p(t_2|t_1)$ に加えて、トピック内単語出現確率 $p(w_1|t_1), p(w_2|t_2)$ を用いて、式 (5) で求めた^{*5}。

$$evocation_score = p(w_1|t_1) * p(t_2|t_1) * p(w_2|t_2) \quad (5)$$

また、評価イメージを図 7 に示す。

トピックモデルでは、それぞれのトピックは単語の出現確率で構成されており、確率の大きな単語ほどそのトピックへの寄与率が高いと考えることができる。よって、確率の大きな単語 (トピックを代表しやすい単語・概念単語) で構成された想起ペアに制限する場合、評価フロー (1) により Top-n を設定する必要がある。なお、従来手法の評価の際には、評価用データが提案手法と同一となるように設計し、想起度合いの算出には単語間 cosine 類似度を用いた。

^{*5} トピック内単語出現確率を用いない場合は、用いる場合よりも平均的に精度が低下したため割愛する。

^{*4} <https://translate.weblio.jp/>

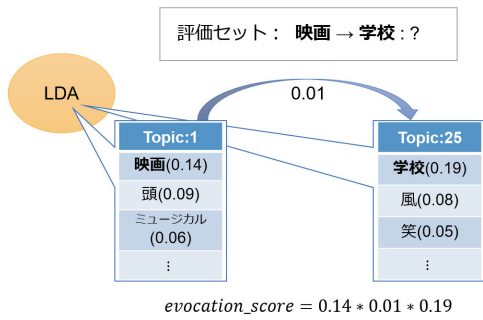


図 7 想起度合いの算出イメージ

Fig. 7 Calculation image of evocation score.

表 4 PWN-Evocation における精度比較

Table 4 Accuracy comparison in PWN-Evocation.

Method	ρ
LSA (従来)	0.090 ± 0.026
TM	0.028 ± 0.028
MCNN	0.027 ± 0.027
LSA&TM	0.125 ± 0.028
LSA&MCNN	0.122 ± 0.027
評価データ数	1,007 ± 66

表 5 Japanese-Evocation における精度比較

Table 5 Accuracy comparison in Japanese-Evocation.

Method	ρ
LSA (従来)	0.174 ± 0.030
TM	0.219 ± 0.008
MCNN	0.192 ± 0.007
LSA&TM	0.293 ± 0.025
LSA&MCNN	0.281 ± 0.022
評価データ数	825 ± 145

4.5 実験結果

実験結果を表 4, 表 5 に示す. 本実験では, LDA のランダム性を考慮し 5 試行の評価値の平均と標準偏差を表示している. そのため, 各試行ごとに評価データ数が変化することに注意されたい. ただし, 各試行における手法間の評価データは同一である. また, トピック数は 200 とし, トピック内上位 3 単語を対象とした. ここでは, 文献 [9] の LSA を用いた共起情報ベースの想起抽出手法を従来手法とし, 提案手法 2 種 (TM-LDA, MCNN-LDA) との比較結果を示している*6. 加えて, 従来手法と提案手法による予測結果のアンサンブルを行った結果を合わせて掲載している. なお, 従来手法では予測出力の値が -1 から 1, 提案手法では 0 から 1 の値に制限されるため, アンサンブル手法ではスコアを最小値 0・最大値 1 の正規化を行った後, 予測出力値の加算平均をとったスコアを新たな予測値とした. さらに図 8 に, J-E において, 閾値 top-n (トピック内上位何単語を対象とするか) を 1 から 100 の範囲で変更

*6 なお, 図表内では TM, MCNN と略記している.

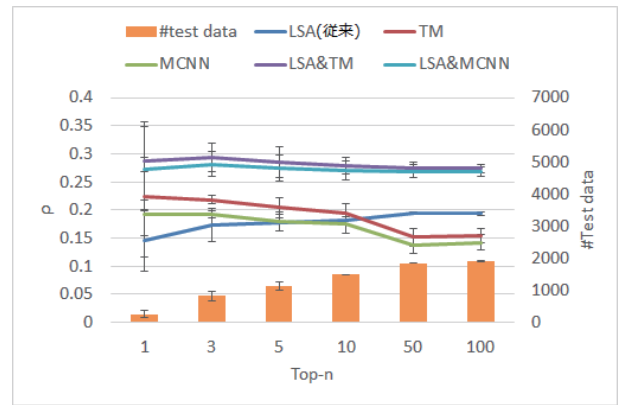


図 8 Top-n の違いによる精度比較 (J-E)

Fig. 8 Accuracy comparison by top-n difference (J-E).

表 6 人手による想起度合いの評価 (J-E)

Table 6 Manual evaluation of evocation score (J-E).

Target	ρ
sub1	0.538
sub2	0.491
sub3	0.469

した際の精度比較結果を示す. また, 表 6 に, J-E における人手による想起度合いの評価結果を示す. これは, ターゲットとなる被験者の想起度合いを, その他 2 名が予測するような形式を想定し, 他 2 名の想起スコアの平均値を用いて, ρ を算出したものである. なお, PWN-E を用いなかった理由は, 被験者情報とそれぞれの想起スコアが紐付けられていないためである. これらの結果より, 確認できることを以下に記述する.

- PWN-E と J-E の比較をすると, 全体的に PWN-E における予測精度が低いことが分かる. これは日本語・日本人のコーパス (不満調査データ) から抽出した想起関係を用いて, 英語・アメリカ人の想起関係を予測しようとしている点がミスマッチとなった可能性があると考えられる. 仮に, PWN-E における想起傾向が J-E と類似しているのであれば, 英日間の精度差は小さくなるであろう. 実際, 図 5, 図 6 より, 想起データのスコア分布が大きく異なっており, PWN-E と J-E とで想起傾向は異なっていると考えられる. 特に提案手法では, 予測精度が 0 に近い値となっており, 英日間での汎用性の面を考えると, 共起ベースの LSA による抽出手法の方が良い結果であることが分かる. また, 日本語の想起データにおいては, 提案手法が既存の手法と同程度以上の精度となっていることが分かる.
- 提案手法間の比較をすると, 全体的に TM-LDA の方が MCNN-LDA より優位であることが分かる. TM-LDA は MCNN-LDA と比べて計算速度の面で圧倒的に高速であり, 演算速度が要求される場合は大きなメリットがある. ただし, MCNN-LDA では入力次元の拡張に

よりメタデータを付与することで、より個々の特徴をとらえた想起関係を得られる可能性もあることは留意しておきたい。

- 従来手法と提案手法をアンサンブル（予測結果の正規化平均を新たな予測値とする）した結果をみると、PWN-E・J-Eの両方で予測精度の向上が見られ、特にJ-Eでは大きな精度向上が確認できる。これは、単語の共起情報に強い従来手法と、想起の方向情報に強い提案手法の双方の良さが組み合わさることで発現した効果であると考えられる。また、提案手法2種の間で比較すると、単体モデルにおける予測精度の優劣関係を引き継いでいることも確認できる。
- 図8より、単体モデル3種とアンサンブルモデル2種では、予測精度の変動傾向が異なることが分かる。特に、従来手法と提案手法2種では、Top-nが10ほどで精度の逆転が起きていることが分かる。これは、提案手法では各トピックをよりよく表現する単語（概念単語）は単語出現確率の大きいもの、つまり、Top-nを小さく制限した場合に抽出されやすくなり、この制限を緩くしていくことでトピックを表現する単語とは関連性の低い単語が抽出されやすくなることで、トピック間遷移確率が有効に働いていない可能性が考えられる。その点に関して、式(5)により、ある程度制御できていると考えられるが、それだけでは不十分であるといえるだろう。

4.6 考察

4.5節の結果では、従来手法と提案手法をアンサンブルしたLSA&TM-LDAが総じて最も精度が高くなることが確認された。これは、単語の共起情報と想起の方向情報を組み合わせることで発現した効果であると考えられる。しかし、提案手法におけるLDAの学習の際には共起情報を利用しており、提案手法のみでも共起情報と方向情報の調和が取れることを期待していた。調和が取れなかった原因として、LDAにおける共起情報は各トピック内で完結しており、トピック間の関係に重きをおいている提案手法では、その情報を取り出すことが難しい状態であったのではないかと考えられる。ただし、同一トピック遷移の場合はこれに該当しない。よって、共起情報もうまく扱えるようにするためには、次投稿予測モデルから想起関係を抽出する際に、トピック内における単語間関係を考慮することや、文書のベクトル表現を、Word2Vec[3]やBERT[5]などの、より表現能力の高い手法に変更することも有用であると考えられるため、検証の必要があるだろう。また表6より、スピーアマンの順位相関係数の理論限界値は1であるが、人手による想起度合いの予測精度は0.5前後と、低い値となっていることが分かる。これは、想起度合いの予測という問題設定の困難さが原因であると考えられ、本実験

による予測精度の妥当性が示唆される。

5. まとめ

本論文では、人間の知覚情報の一種である想起に注目し、ソーシャルメディアを用いた想起関係の自動抽出手法の開発を試みた。従来の手法では、想起関係の抽出に単語の共起情報を用いており、想起の方向性を表現できていないという問題があると考え、本論文では、投稿間想起仮説に基づく次投稿予測モデルを用いた想起抽出手法を提案した。さらに、実験により、英日2種の想起データを用いて、実際に抽出された想起関係度合いの評価を行い、特に日本語データにおいて、従来手法と比較して提案手法の予測精度が同程度以上となることが確認された。また、改良モデルとして両手法による予測結果をアンサンブルした場合、英日の想起データともに予測精度が向上することが確認された。これにより、想起抽出モデルに単語の共起情報だけではなく、想起の方向情報を取り入れることの重要性を示唆し、より精度の高い想起抽出モデルが構築できることを示した。今後は、想起抽出モデルに共起情報と方向情報を適度に組み込んだ単一モデルを開発し、より予測精度の高いモデルを追求していく。

謝辞 本論文では、株式会社 Insight Tech が国立情報学研究所の協力により研究目的で提供している「不満調査データセット」を利用した。

参考文献

- [1] Dencik, L., Hintz, A. and Carey, Z.: Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom, *New Media & Society*, Vol.20, No.4, pp.1433-1450 (online), DOI: 10.1177/1461444817697722 (2018).
- [2] Das, S., Behera, R.K., kumar, M. and Rath, S.K.: Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction, *Procedia Computer Science*, Vol.132, pp.956-964 (2018).
- [3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26*, Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger, K.Q. (Eds.), pp.3111-3119, Curran Associates, Inc. (2013) (online), available from (<http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>).
- [4] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L.: Deep contextualized word representations (2018).
- [5] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018).
- [6] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding (2019).
- [7] Silberer, C., Ferrari, V. and Lapata, M.: Models of

Semantic Representation with Visual Attributes, *Proc. 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.572-582, Association for Computational Linguistics (2013) (online), available from (<https://www.aclweb.org/anthology/P13-1056>).

[8] Bruni, E., Tran, N.K. and Baroni, M.: Multimodal distributional semantics, *Journal of Artificial Intelligence Research*, Vol.49, p.147 (2014).

[9] Boyd-Graber, J., Fellbaum, C., Osherson, D. and Schapire, R.: Adding Dense, Weighted Connections to WORDNET, *Proc. 3rd International WordNet Conference*, pp.29-36 (2006).

[10] Grice, G.R.: Stimulus intensity and response evocation, *Psychological Review*, Vol.75, No.5, pp.359-373 (1968).

[11] Buss, D.M.: Selection, evocation, and manipulation, *Journal of Personality and Social Psychology*, Vol.53, No.6, pp.1214-1221 (1987).

[12] Larsen, R.J. and Buss, D.M.: *Personality psychology: Domains of knowledge about human nature*, McGraw-Hill (2002).

[13] Hayashi, Y.: Predicting the Evocation Relation between Lexicalized Concepts, *Proc. COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, The COLING 2016 Organizing Committee, pp.1657-1668 (online), available from (<https://www.aclweb.org/anthology/C16-1156>) (2016).

[14] Cramer, I.: How Well Do Semantic Relatedness Measures Perform?: A Meta-study, *Proc. 2008 Conference on Semantics in Text Processing, STEP '08*, pp.59-70, Association for Computational Linguistics (online), available from (<http://dl.acm.org/citation.cfm?id=1626481.1626487>) (2008).

[15] Ma, X.: Analyzing and propagating a semantic link based on free word association, *Language Resources and Evaluation*, pp.819-837 (2013).

[16] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R.: Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, Vol.41, pp.391-407 (1990).

[17] Moss, H. and Older, L.: *Birkbeck word association norms*, Psychology Press (1996).

[18] Lindzey, G. and Runyan, W.M.: *A History of Psychology in Autobiography, Volume IX*, Clark University Press (1936).

[19] Fellbaum, C.: WordNet An Electronic Lexical Database, *The MIT Press*, pp.29-36 (1998).

[20] Lesk, M.: Automatic sense disambiguation using machine-readable dictionaries, *Proc. SIGDOC*, pp.391-407 (1986).

[21] Leacock, C. and Chodorow, M.: *WordNet: An Electronic Lexical Database*, MIT Press (1998).

[22] Nikolova, S., Boyd-Graber, J. and Fellbaum, C.: *Collecting Semantic Similarity Ratings to Connect Concepts in Assistive Communication Tools*, pp.81-93, Springer Berlin Heidelberg (2012).

[23] Nelson, D.L., McEvoy, C.L. and Schreiber, T.A.: The University of South Florida free association, rhyme, and word fragment norms, *Behavior Research Methods, Instruments, & Computers*, Vol.36, No.3, pp.402-407 (2004).

[24] Mitsuzawa, K., Tauchi, M., Domoulin, M., Nakashima, M. and Mizumoto, T.: FKC Corpus: A Japanese Corpus from New Opinion Survey Service, *Proc. Novel Incentives for Collecting Data and Annotation from People:*

Types, Implementation, Tasking Requirements, Workflow and Results (2016).

[25] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent dirichlet allocation, *Journal of Machine Learning Research*, Vol.3, pp.993-1022 (2003).

[26] Wang, Y., Agichtein, E. and Benzi, M.: TM-LDA: Efficient Online Modeling of Latent Topic Transitions in Social Media, *Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, pp.123-131, ACM (online), DOI: 10.1145/2339530.2339552 (2012).



内田 脩斗 (学生会員)

2018年3月名古屋大学工学部電気電子・情報工学科卒業。同年4月同大学大学院工学研究科博士課程前期課程情報・通信工学専攻に入学、現在に至る。主として自然言語処理に関する研究に従事。人工知能学会、言語処理学

会会員。



吉川 大弘 (正会員)

1997年名古屋大学大学院博士課程修了。同年カリフォルニア大学バークレー校ソフトコンピューティング研究所客員研究員。1998年三重大学工学部助手。2005年名古屋大学大学院工学研究科 COE 特任准教授。2006年10月同研究科准教授、2020年4月鈴鹿医療科学大学医用工学部医用情報工学科教授、現在に至る。主としてソフトコンピューティングとその応用に関する研究に従事。博士(工学)。IEEE, 人工知能学会, 日本知能情報ファジィ学会, 進化計算学会各会員。



古橋 武

1985年名古屋大学大学院工学研究科博士後期課程電気系専攻修了。工学博士。2004年名古屋大学大学院工学研究科計算理工学専攻教授、2020年名古屋大学国際機構特任教授、名誉教授。ソフトコンピューティング、感性工学に関する研究に従事。1996年日本ファジィ学会論文賞受賞。IEEE, 日本知能情報ファジィ学会名誉会員。