

隠れた良作を推薦可能な Web 小説レコメンドシステムの提案

Recommending Hidden Masterpieces: Towards an Attentive Web Novel Recommender System

小坂 直輝^{1*} 小林 哲則¹ 林 良彦¹
Naoki Kosaka¹ Tetsunori Kobayashi¹ Yoshihiko Hayashi¹

¹ 早稲田大学 理工学術院

¹ School of Science and Engineering, Waseda University

Abstract: Websites accommodating user-generated content have been becoming popular among the general users who would like to share and enjoy the published works. An effective and efficient search/recommender system is however required to adequately link the publishing users and the consumer users. In the present study, we target a Website that provides an arena for sharing Japanese novels authored by general users and propose a system for making a personalized recommendation even for yet to popularized works based on the quality of the sampled sentences. This may contribute to discovering “hidden masterpieces.” The proposed system may also be able to uncover works that might match a user’s preference by measuring the similarity to already-read works. This paper specifically presents machine learning-based frameworks for estimating sentence qualities and predicting inter-works similarities. The preliminary experimental results suggest that the proposed frameworks are promising in implementing a recommender system that may satisfy the desiderata.

1 はじめに

1.1 研究背景と目的

近年、インターネット上で個人が小説を投稿できる、小説投稿サイトが多数開設され、数多くの作品が投稿されている。こうしたサイトに投稿される作品は一般に web 小説やオンライン小説と呼ばれ、これらの作品の中には商業用に書籍化・映像化される作品も存在している。

一方でこれら web 小説に関して、その作品数の多さから読者が好みの作品を探すのが難しいという問題や、人気の作品や最新の作品といった限られた少数の作品に大多数の読者が集中するという問題がある。

こうした問題は特に、有名な作品は一度は見かけたことがある、一定以上の利用経験がある読者にとっては重大であり、そういった読者が大量にあるマイナーな作品から、次に読むための良い作品を探そうとすると多大な

労力を要求される。結果として誰にも読まれない良い作品、いわゆる隠れた良作ができてしまう。

書き手の立場からすればこれは、作品を読んでもらうことの困難さに繋がり、良い作品を書いてもその作品が読者の目に入らない可能性があると言える。さらにこうした問題は読者、筆者両方について小説投稿サイトの利用を止めたり、ひいては web 小説そのものから離れる要因になることも考えられるので、解決する必要がある。

本研究の目的は、上記問題点を解決し、ユーザが作品を探すのを助けるレコメンドシステムを実現することである。具体的には、日本最大級の小説投稿サイトであり、約 70 万の作品と 160 万人のユーザ数を有する「小説家になろう」*1の作品を対象に、作品の本文とジャンルやキーワードといった付属情報から、作品の類似度や質を推定するモデルを構築することで、読者が付与した情報の無い隠れた良作の推薦や、ユーザによる推薦基準の操作が可能なレコメンドシステムを提案する。

特に本稿では、類似度や質を定める有効な情報について検討し、機械学習を用いた推定実験の結果を議論する。

* 連絡先：早稲田大学理工学術院

〒 162-0042 東京都新宿区早稲田町 27 早稲田大学 40 号館 701 号室
E-mail: kosaka@pcl.cs.waseda.ac.jp

*1 <https://syosetu.com/>

1.2 関連研究

一般にアイテムベース協調フィルタリングなどの代表的な推薦アルゴリズムは、新規アイテムはユーザの評価履歴が無いため推薦できない、コールドスタート問題を抱えている [1][2][3]。本研究における隠れた良作とは評価がついていれば高い評価となるであろうが、評価が無いためにユーザの目に入らない作品であり、この点でコールドスタート問題と関連がある。コールドスタート問題への対処法として、アイテム自身の特徴を利用する内容ベースの推薦アルゴリズムならユーザの評価履歴が無くとも推薦が可能であり [1][2][3]、本研究でも web 小説のレコメンドにおいて作品自体の情報を利用する。

また、文章を何らかの側面から評価するという試みについても様々な研究がある。例えば日本語の文章の読みやすさを評価する試みとして、[4]では、文の長さや各文字種の頻度から読みやすさを求める評価式が提案されていて、[5]では文章の難易度の基準となるコーパスを定め難易度ごとの言語モデルを作製し、文章の難易度を判定する手法が提案されている。本研究は読みやすさに限らない総合的な文章の質を、機械学習を用いて、様々な作品情報から予測する。

さらに [6]ではトピックモデルや、センチメント分析によるプロットライン判定、文章の文体の分析によりベストセラーを予測するという試みが行われている。高い評価を得る作品には潜在的に共通するパターンが存在すると考え、そういった作品をコンピュータに予測させるという点で本研究と関連がある。

web 小説のレコメンドに関連する研究としては、お気に入り登録のリンク構造を用いる [7] や、作品のあらすじを用いる [8]、文体の類似度を考慮する [9] があるが、本研究は作品の本文も用いるという点で [7][8] と、類似度以外の観点も考慮するという点で [9] と差異がある。また、評価者とシステム利用者の評価基準の不一致に着目した研究 [10] も存在する。

2 提案システム

2.1 読者が小説を選び読む際の観点

提案するシステムはレコメンドシステムであるので、当然ながら提示する作品はユーザが読みたいと思う作品や読んで良かったと感じる作品であるのが理想的である。そこでまず読者が小説を読むか決めるポイントや、読んで良かった、面白かったと感じる要因について考え、表 1 に示すように主に三つの観点に分けてまとめた。

一つ目は小説の種類で、その小説がどのジャンルのどういった形式で書かれたものかといった観点である。この観点は、例えば SF が好きで SF しか読まない読者がいたり、女性が女性向け作品を読んだりというように、小説を選ぶ際に大きく影響し、また、好みのジャンルの方が読んで良かったと感じやすいといったように読書中・読後の感じ方にも影響すると考えた。この 1 の観点は、好みなどによる個人差が大きく、一方で機械による判断は比較的容易だと考えた。

二つ目は、ストーリーやキャラクター、世界観といった小説の内容であり、一般に小説の面白さといった部分に大きく関わるであろう観点である。この 2 の観点は、1 の観点程ではないにせよ好みによる個人差が大きく、一方でこの観点においてどれだけ優れているかを機械によって判断するのは非常に難しいと考えた。

三つ目は、読みやすさや表現の豊富さといった小説の文章に関わる観点であり、これも読書中・読後の感じ方に影響すると考えた。特に web 小説においては、編集者の手が入り一定以上の文章の質が保証される商業書籍と異なり、読みやすさや文法の誤り、誤字脱字といった部分でも、優れた作品からそうでないものまで幅があるので、この 3 の観点も重要である。そしてこの観点は、好みなどによる個人差は比較的小さく、機械による判断もある程度可能なのではと考えた。

表1 読者が小説を選び読む際の観点

1	小説の種類 (ジャンル, 形式, 男性向け女性向け 等) 好みによる個人差: 大 機械による判断: 易
2	小説の内容 (ストーリー, キャラクター, 設定, 世界観 等) 好みによる個人差: 中~大 機械による判断: 難
3	小説の文章 (読みやすさ, 表現の豊富さ, 描写の丁寧さ, 文法 等) 好みによる個人差: 小 機械による判断: 普~難

本研究で提案するレコメンドシステムでは推薦時に S スコアと Q スコアという二つの指標を用いるが、前者で主に 1 の観点に、後者で主に 3 の観点に対応することで、ユーザが読みたいと思うような作品や読んで良かったと感じる作品を推薦する。各指標については後の 2.3 節や 4, 5 節で詳しく述べる。また提案するシステムにおいて、2 の観点に関しては機械による判断の難しさなどから現段階で直接的に対応する機能は考えていない。

2.2 現状の検索システムの問題点

レコメンドシステムを提案するにあたり、小説投稿サイトにおける現状の作品検索システムがどういったもの

であるかその問題点と共に説明する。小説投稿サイトにおける作品検索システムの概要図を図 1 に示す。ユーザが作品を探す流れは、まずジャンルやキーワードといった検索条件を指定し、次に検索結果を表示する順序を指定すると、作品が指定した順に並べて複数ページに渡り表示されるといった流れである。

ここで、この作品の並べ方には複数の選択肢があるが、基本的には人気の作品や最近の作品が前に表示される。そのため小数の人気の作品がユーザの目に入りやすく、そういった作品に読者が偏るといったことが起こる。これはとりあえず人気の作品や最近の作品が読みたいというような、ライトユーザには適している大きな問題はない。一方で、多くの作品を読んでいるユーザが、後ろに表示されている大量のマイナーな作品から良い作品を探そうとすると多大な労力を要求されるので、そういったユーザにとっては問題であり、より適した別の作品提示方法も必要だと考えられる。

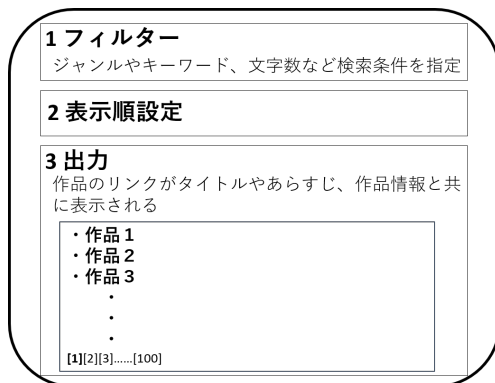


図1 現状の検索システムの概要図

また、別の問題として作品を並べて表示する際の人気度を示す指標が累積値のようなもので、作品を比較する指標として適切でないという点がある。例えば「小説家になろう」*1において作品を人気順に表示するために用いられている指標の一つに総合評価点があるが、これはユーザが着けた作品の評価やブックマークの合計値から計算した値であり、一桁の作品から数十万点の作品まで存在する、上限のない累積値である。累積値では10人が10点と評価した作品よりも100人が5点と評価した作品の方が値が大きいといったことが起こりえるので、作品を比較する指標として適切とは言えない。

小説投稿サイトには「小説家になろう」の他にも「カクヨム」*2や「エブリスタ」*3など多くのものがあるが、作品検索システムの形は類似していて、上記問題点は複数

*2 <https://kakuyomu.jp/>

*3 <https://estar.jp/>

の小説投稿サイトにおいてある程度共通する問題である。

2.3 レコメンドシステムの設計方針

本研究の目的や既存の検索システムの問題点を踏まえ、提案するレコメンドシステムの要件を次の二つにまとめた。一つ目が読者が少ない、いわゆるマイナーな作品もレコメンドの提示候補となりうることで、二つ目がレコメンドにおいて何を重視するかという、レコメンドの基準をユーザが操作できることである。

以上二つの要件を満たすため、提案するレコメンドシステムでは、SスコアとQスコアという二つの指標を用いる。Sスコアはある作品が他の作品とどれだけ似ているかを示す指標で、これによりユーザの過去の読書傾向に似た作品を提示する。一方Qスコアは、主に表1の小説の文章の観点において、作品の質がどれだけ良いかを示すもので、これによりユーザがより満足できる作品を提示する。

ここで、上記のように作品の類似度や質を示すものとして定めた二つの指標、SスコアやQスコアについて、その値をどう求めるかが重要である。単純な方法として、作品の類似度や質を読者の一致率や評価の高さと見なし、読者が付与したデータから求める方法が考えられるが、それでは、読者が付与したデータの少ないマイナーな作品には各指標が適切に求められない可能性がある。そこでSスコアとQスコアのそれぞれにおいて、小説本文も含めた作品自体の情報から各指標を予測するモデルを構築し、これを用いることで読者が付与したデータのないマイナーな作品についてもレコメンドを可能にする。

また、最終的なレコメンドにおいては、SスコアとQスコア、二つの指標を合わせた値の高さで作品の提示を行うが、その際、各指標の重みを調整可能にし、レコメンドの基準をユーザが操作できるようにする。例えばSスコアの重みを大きくすれば、提示する作品がより過去の読書傾向との類似性を重視したものになる。この仕組みから、より各ユーザに適した形で、ユーザが作品を探すのを助けるシステムになると考えた。

3 データ

本研究では「小説家になろう」*1からAPIやスクレイピングを用いて取得した小説のデータを利用する。まず、本研究で取得するデータに含まれる情報を、その種類により分類したものを表2に示す。表2における読者が付与した情報は作品ごとにその有無や量が異なるが、作品自体の情報は基本的に全ての作品に共通して存在する。

レコメンドに用いる S スコアと Q スコアという二つの指標をモデルから予測させる際は、表 2 における作品自体の情報を入力として用いることで、マイナーな作品についても対応する。

表2 取得データに含まれる情報の分類

読者が付与した情報	評価点, 感想, レビュー 等
作品自体の情報	
作者による付属情報	ジャンル, 文字数 等
作品の中身	本文, 文章の品詞傾向 等

以下具体的なデータの取得状況を説明する。まず、「小説家になろう」で公開されている全作品を対象に、提供されている API を用いて取得できる 33 項目の作品情報を取得した。ここにはタイトルやジャンル、文字数やユーザの評価点などが含まれ、670,834 作品分の作品情報を取得した。このうち、368,576 作品が長編小説、長編小説の 85,649 作品が完結済であった。ここで長編小説とは、作者が小説を一部分ずつ投稿、更新していくタイプの作品であり、「小説家になろう」の作品はこの長編小説と、短い作品を一度に投稿する短編小説の二種類に大別される。本研究では、ユーザの需要やデータ数の観点から、5 万字以上の長編小説もしくは 1 万字以上で完結済の条件を満たす 116,529 作品を扱うこととした。

次に、上記 116,529 作品を対象に小説の本文を取得した。具体的にはスクレイピングにより、小説のランダムな三か所から 1,200 文字弱ずつの文章を取得した。このような取得の仕方をしたのは、全対象作品の全文章を取得するのはデータの規模から容易ではない一方、作品の先頭など特定の一か所だけから取得すると、その作品全体の文章傾向とずれ偏ったものとなる恐れがあるからである。

最後に、同じく上記条件を満たす長編小説を対象に、作品の文章評価とストーリー評価を取得した。「小説家になろう」には読者が作品の文章とストーリーを各 1~5 点で評価できる仕組みが存在する。API から取得できるのは全評価点の合計のみで、文章評価とストーリー評価の割合は考慮出来ないが、作品ページにはそれぞれの評価の合計が公開されているのでこれをスクレイピングにより取得した。

データの取得には時間がかかり、その間に削除された作品や何らかのトラブルで取得が出来なかった作品もあるため、実際に取得したのは本文が 114,233 作品分、文章評価とストーリー評価が 112,851 作品分である。

その他の取得は行っていないが利用を検討しているデータとして、各作品に書かれている感想や、ユーザの評価履歴といったデータがある。

4 作品の質予測

ここからはレコメンドに用いる指標であり、作品の類似度や質を示すものとして定めた二つの指標、S スコアと Q スコアについて作品自体の情報から予測が可能かを調べるために行った、分析や実験の内容と結果について述べる。まずこの 4 節では作品の文章の質などを示す Q スコアについて触れる。

4.1 Q スコアの設定

Q スコアを定めるにあたり、この指標は作品の質の高さを示すもので、かつ作品の比較を行うのに適した指標であることが望ましい。そこで今回、式 1 で求められる指標を Q スコアとして設定した。

$$16 \times \frac{\text{累積文章評価}}{\text{評価人数}} + 10 \times \frac{\log_2(1 + \text{レビュー数})}{\sqrt[4]{\text{評価人数}}} \quad (1)$$

取得したデータの中で、Q スコアを求める参考になる情報に文章評価があり、これは累積値であるが、取得したデータには評価人数も含まれているので、累積文章評価を評価人数で割ることで、比較が可能なる 5 点を最大とする平均文章評価点を求められる。また、レビューというのは「小説家になろう」*1 においては読者が作品を推薦する目的で書かれるもので、多いほど良い作品と判断できる。これらの観点から今回式 1 で求められる値を Q スコアとして設定した。なお、対数や 4 乗根、定数倍はスケール調整のためのものであり、Q スコアの値の約 8 割が平均文章評価点により定まるようになっている。これはレビューが書かれていない作品も多くあることを考慮した結果である。

文章評価を取得した作品の内、評価人数が 10 人以上の作品に対し、上式 (1) を用いて Q スコアを計算した結果のヒストグラムを図 2 に示す。縦軸が作品数、横軸が Q スコアで、最大値が 97.8、最小値が 35.2 と一定の範囲に取り、正規分布に近い分布となっているので、作品を比較する指標としてある程度適切だと考えた。

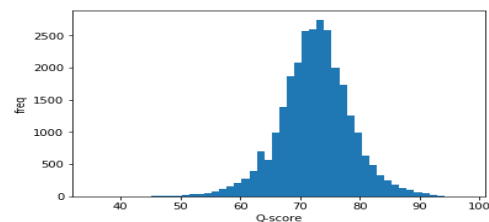


図2 Q スコアの分布

4.2 Q スコア予測実験

4.2.1 事前分析

まず、事前の分析として、取得できるデータに含まれる、作品の総文字数、会話率、完結済かの三つの特徴量について Q スコアとの相関を調べたが、明確な傾向は確認出来なかった。例として文字数と Q スコアの散布図を図 3 に示す。

文字数と会話率のどちらにおいても Q スコアとの明確な相関は見られず、また完結済かにおいても、微かに完結済の作品の方が平均 Q スコアが高かったが有意な差は確認できなかった。

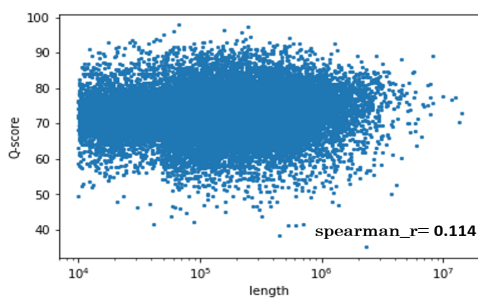


図3 文字数と Q スコアの相関

4.2.2 機械学習モデルによる予測

まず、事前分析で用いた三つの特徴量について、これを合わせて用いることで予測ができないかと考え重回帰分析とシンプルな Neural Network モデルによる学習と予測を行ったが、精度は悪かった。(結果は次の 4.2.3 節に示す)

そこで、使用する作品の特徴量を増やし改めて Neural Network モデルを構築し、学習・評価を行った。構築したモデルの構造を図 4 に示す。

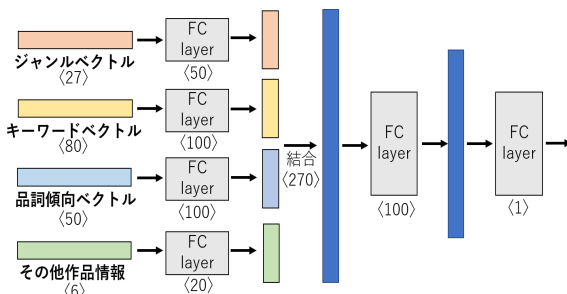


図4 Q スコア予測モデル

入力ベクトルと含まれる特徴について、まず、ジャンルベクトルは「小説家になろう」におけるジャンル分類

において、作品がどのジャンルに含まれているかを表している。キーワードベクトルは作者が作品に設定するキーワードの内、今回用いたデータセットにおいて出現回数が一定数以上だったキーワードにおいて、各キーワードが含まれているかを表したバイナリベクトルである。品詞傾向ベクトルは作品の本文を形態素解析し、各品詞の出現頻度を表したベクトルである。形態素解析には日本語自然言語処理ライブラリの GiNZA^{*4}を用いた。その他の作品情報としては、事前の分析で用いた三つの特徴量に加え、タイトルの長さ、漢字の割合、文密度(文字数/行数)を用いた。

また、ここまでは Q スコア予測モデルの入力として用いる作品のベクトルを取得したデータから直接生成していたが、文書のベクトル化手法である Doc2Vec[11] を用いて、作品の本文から教師なし学習を行い生成した作品のベクトルを、入力として用いる場合についても実験を行った。Doc2Vec を用いるにあたっては、GiNZA を用いて分かち書きした単語全てを用いる場合と、助詞や助動詞といった一部品詞を除いた場合、[11] で紹介されている二つの手法である dmpv と DBow のそれぞれを用いた場合、[12] で紹介されている最適パラメータを参考にだまかにパラメータを変化させた場合など様々な組み合わせを試した。

また、図 4 における入力ベクトルと Doc2Vec によって生成した作品ベクトルを合わせて Neural Network の入力として用いる場合や、それぞれを単独に用いたモデルの出力をアンサンブルした場合についても実験を行った。

4.2.3 実験結果と考察

今回行った機械学習モデルによる Q スコア予測実験の結果を、学習において最小化を目指した平均二乗誤差と、モデルの精度を考える目安となるスピアマンの相関係数と共に表 3 に示す。Q スコアの値の大小が正しく予測出来ていることが望ましいので、スピアマンの相関係数が 1 に近いほど良いモデルだと言える。表 3 における MLP (Doc2Vec のみ) と記したものが Doc2Vec による作品ベクトルから Neural Network モデルによる予測を行ったもので、MLP (Doc2Vec 併用) が図 4 の入力ベクトルと Doc2Vec による作品ベクトルを共に Neural Network の入力としたもの、MLP (アンサンブル) がそれぞれ単独に用いた予測値を 7:3 で重み付き平均した場合である。

なお Doc2Vec については複数試した組み合わせの中で最も精度が良かった、前処理で一部品詞を除き、dmpv

^{*4} <https://megagonlabs.github.io/ginza/>

のモデルで 35 エポック学習させたものを用いた結果について掲載している。実験では評価人数が 10 人以上の作品で本文の取得が完了している 26,859 作品を利用し、そのうち 8 割を訓練データ、訓練データの 1 割を validation 用として利用した。

表3 Q スコア予測精度

モデル	特徴の種類数	誤差	相関
重回帰分析	3	34.35	0.126
MLP	3	34.29	0.143
MLP	9	27.02	0.468
MLP (Doc2Vec のみ)	1	30.64	0.363
MLP (Doc2Vec 併用)	10	28.27	0.436
MLP (アンサンブル)	10	26.43	0.487
訓練データの平均を出力		34.66	—

事前の分析で用いた、文字数、会話率、完結済かの三つの特徴量だけでは、Q スコアの予測精度は、訓練データにおける Q スコアの平均を出力した場合と大差ない。一方で、ジャンル、キーワード、品詞傾向、タイトルの長さ、漢字の割合と文密度というさらに六種類の特徴を用いたモデルでは、予測精度に改善が見られた。また Doc2Vec による作品ベクトルのみを用いた場合でも多少は予測が行えているが、その精度は図 4 の九種類の特徴を用いたモデルより悪く、九種類の特徴と併用した場合でも同様だった。だが、九種類の特徴と Doc2Vec による作品ベクトルを個別に用いた予測値をアンサンブルした場合は精度が上がっているため、パラメータ調整やモデルの構造の検討が十分でない可能性も考えられる。

次に正解値と予測値の関係を確認するため、最も精度が良かった場合における正解値と予測値の散布図を図 5 に示す。

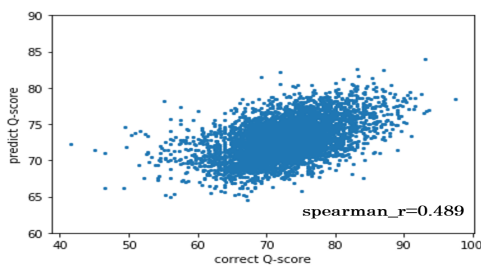


図5 正解値と予測値の相関

図 5 と表 3 の結果から正解値と予測値にある程度相関があることが確認できた。また図 5 から、正解値に対し予測値の分散が小さく、学習データが少ない Q スコアが 65 以下や 85 以上などの作品に対し正確に予測できず予測誤差が下がらなかったのだと考える。今後精度を改

善する方法として、一つにはさらに別の特徴を用いることが考えられる。今回利用出来なかった特徴として、様々な自然言語処理タスクで成果を上げている BERT[13] を用いて求めた文章の分散表現があり、これを用いることで予測精度が向上する可能性がある。また、各特徴量の予測値に対する影響度を調べたり、それに応じてモデルの構造やパラメータを見直すことでも、精度を向上させられると考える。

さらに、今回設定した Q スコア自体についても、ユーザの評価履歴や感想といったデータの取得後、改めて適切な指標設定を検討する余地がある。

5 作品の類似度予測

この 5 節では作品の類似度を示す S スコアについてその求め方や、妥当性を確認するために行った実験について述べる。

5.1 S スコアの設定

S スコアを求めるにあたり、作品の類似度の人手による正解データは存在せず、類似度の正解データを新たに作成するのも困難である。一般にアイテムベースの協調フィルタリングや、二つの文の類似度を求めるタスクである Semantic Textual Similarity では、それぞれのアイテムや文の多次元ベクトルによる表現を求め、そこからコサイン類似度やユークリッド距離、Jaccard 係数などを計算することで類似度を求めるという手法が多く使われる。同様に作品の多次元ベクトルを求める方法として、同じユーザが同じ評価をしている作品は似ているという観点から、ユーザの評価履歴を用いる方法が考えられるが、「小説家になろう」*¹のユーザ数は 160 万人に上り、その一部にしても取得には膨大な時間がかかるためこの方法は現実的でない。またデータの偏りやスパース性といった問題もある。

そこで今回、4 節でも用いた Doc2Vec による作品ベクトルを利用し、この作品ベクトル同士のコサイン類似度を求めこれを S スコアとして利用することとした。「小説家になろう」において作品ジャンルやキーワードは作者が自由に設定できるが、こうして求めた S スコアにより、設定されたジャンルやキーワードは異なるが実際は似ている作品も探せると考えている。

5.2 S スコア予測実験の結果と考察

S スコアに関する実験においては本文の取得が完了している作品のうち、ジャンルがノンジャンルではない

87,237 作品を用いた。ここでノンジャンルの作品とは「小説家になろう」で 2016 年に行われたジャンル分類の変更後にジャンルが再設定されていない作品である。ノンジャンルの作品を除いたのはこの後で述べる S スコアの予測実験や結果の分析で作品のジャンルデータを利用するためである。

S スコア自体は Doc2Vec により生成した作品ベクトルのコサイン類似度を計算することで簡単に求められるが、作品の類似度の正解データが存在しないため、求めた S スコアがどの程度適切なものか直接的には評価が出来ない。そこで今回は作品のクラスタリングを行った結果の確認や、同ジャンル間と別ジャンル間での差異の確認、S スコアを元に同一作品の判定を行う実験などを通して、ある程度の妥当性を確認した。

まず、Doc2Vec により生成した作品ベクトルから Ward 法による作品のクラスタリングを行った結果のデンドログラムを図 6 に示す。ここで S スコアの妥当性を確認するという観点から、S スコアが高い作品同士が近いクラスタとなるよう、各作品ベクトルを正規化してからクラスタリングを行った。Ward 法ではクラスタ間の距離を求める過程でユークリッド距離を利用するが、正規化を行うことでユークリッド距離が小さい作品と cos 類似度が高い作品が対応するためである。

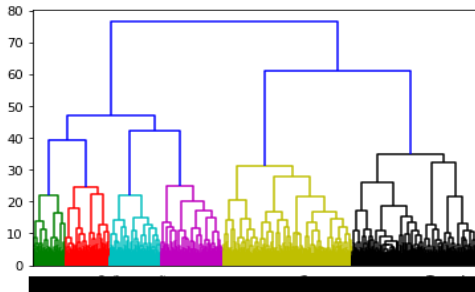


図6 Ward 法による作品のクラスタリング

また図 6 だけではクラスタリングが適切に行われているか判断できないため tf をクラスタ内単語出現頻度、idf を逆作品頻度として tf-idf を求め各クラスタの特徴語抽出を行った結果を図 7 に示す。加えて「小説家になろう」における大ジャンル五つにおいて、各ジャンルがどのクラスタに多く含まれているか求めた結果を表 4 に示す。ここから同一クラスタ内や隣接するクラスタに含まれる作品にある程度類似性があることが確認できた。

例えば図 7 からクラスタ 2 は「情報」や「通信」といった科学と関連するような特徴語が、クラスタ 5 には「好き」や「一緒」といった恋愛と関連するような特徴語が含まれていて、実際に表 4 からクラスタ 2 の 21% が SF ジャンル、クラスタ 5 の 30% が恋愛ジャンルであるこ

クラスタ1 6493作品 軍,兵,者,国,王国,王,兵士,貴族,敵,戦う
クラスタ2 8740作品 よる,情報,行,部隊,通信,敵,軍,可能,確認,現在
クラスタ3 9971作品 攻撃,地面,剣,放つ,力,戦う,一撃,体,構える,光
クラスタ4 12277作品 スキル,冒険者,ギルド,魔法,倒す,レベル,魔物,魔力,ステータス,攻撃
クラスタ5 25392作品 学校,今日,好き,いつも,家,高校,電話,一緒,友達,教室
クラスタ6 24364作品 瞳,笑う,優しい,姿,見詰める,首,部屋,心,体,表情

図7 各クラスタの特徴語

表4 主要ジャンルとそのジャンルを含む割合の多いクラスタ

恋愛	クラスタ 5: 30%	クラスタ 6: 29%
ファンタジー	クラスタ 3: 68%	クラスタ 1: 64%
文芸	クラスタ 5: 35%	クラスタ 2: 28%
SF	クラスタ 2: 21%	クラスタ 4: 10%
その他	クラスタ 2: 8%	クラスタ 5: 5%

とが確認できる。また図 6 と図 7、表 4 から合わせて判断することでクラスタ 1 から 4 にファンタジー、クラスタ 5 と 6 に恋愛系の作品が多く集まっていると判断でき、これはこの二つのジャンルが「小説家になろう」において人気で作品数が多いことを考慮すると妥当な結果だと考えられる。

次に「小説家になろう」の小ジャンル 20 分類について、同ジャンルの作品間と別ジャンルの作品間で求めた S スコアの平均について表 5 に示す。表 5 から同ジャンルの作品間で求めた S スコアの方が高い傾向があることが確認できる。

表5 同ジャンル間と別ジャンル間の平均 S スコア

同ジャンル間の平均 S スコア	別ジャンル間の平均 S スコア
0.197	0.137

最後に S スコアを用いて同一作品判定を行った結果の精度を表 6 に示す。同一作品判定においてはまず、取得した各作品の文章を半分ずつに分け、それぞれ別々に Doc2Vec のモデルを学習させ作品ベクトルを求めた。その後異なる作品の文章から求めたベクトルと、同じ作品の別の場所の文章から求めたベクトルを同量ずつ用意し、それぞれの S スコアを計算して、中央値以上なら同一作品と判定した。

別の場所の文章から求めたベクトル同士であっても、それが同じ作品のベクトルなら S スコアは高くなるべきで、実際にある程度同一作品判定が行えていることが確認できた。

表6 Sスコアによる同一作品判定

判定方法	同一作品判定正答率
Sスコアを利用	0.866
ランダム	0.501

以上いくつかの実験や分析を通して、今回設定し、実際に作品の本文から求めた S スコアが、作品の類似度を示すものとしてある程度妥当なものだと確認できた。今後より適切に S スコアを求める方法として、今回用いた Doc2Vec による作品ベクトル以外の特徴を用いることが考えられる。例えば 4.2.3 節でも触れた BERT[13] を用いて作品の文章の分散表現を求め、そこから同じくコサイン類似度を計算することで S スコアを求める方法なども検討している。

6 おわりに

隠れた良作の推薦や、ユーザによる推薦基準の操作が可能な web 小説レコメンドシステムとして、作品の類似度を示す S スコアと作品の質を示す Q スコアの二つの指標を用いて、この二つの指標を作品自体の情報から予測することで推薦を行うシステムを提案した。加えて、実際にそれぞれの指標を設定し機械学習モデルによる予測実験を行うことで、各指標の妥当性や予測可能性を確認しレコメンドシステムが作製できる見通しが立った。今後は今回用いたモデルをベースとしてシステムを作製しながら、モデルを改善し各指標の予測精度の向上を目指す。平行してデータの追加収集を行い、再実験の実施や、今回用いてないデータの利用も検討する。また、最終的な評価は完成したシステムをユーザに使ってもらい行う予定だが、S スコアと Q スコア、各指標の予測モデルについてもより適切な評価方法がないか考えていきたい。

参考文献

- [1] Xiaoyuan Su, Taghi M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques”, *Advances in artificial intelligence*, 2009.
- [2] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock, “Methods and Metrics for Cold-Start Recommendations”, *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.253-260, 2002.
- [3] 神瀧敏弘, “推薦システムのアルゴリズム (2)”, *人工知能学会誌*, Vol.23, No.1, pp.89-103, 2008.
- [4] 建石由佳, 小野芳彦, 山田尚勇, “日本文の読みやすさの評価式”, *情報処理学会研究報告ヒューマンコンピュータインタラクション 25(1988-HI-018)*, pp.1-8, 1988.
- [5] 近藤陽介, 松吉俊, 佐藤理史, “教科書コーパスを用いた日本語テキストの難易度推定”, *自然言語処理学会第 14 回年次大会発表論文集*, pp.1113-1116, 2008.
- [6] ジョディ・アーチャー, マシュー・ジョッカーズ, 川添節子訳, “ベストセラーコード”, 日本 BP 社, 2017.
- [7] 清水一憲, 伊東栄典, 廣川佐千男, “集合知に基づくオンライン小説のランキング手法の提案と評価”, *情報処理学会研究報告*, 2013.
- [8] 飯田委哉, 伊東栄典, “セレンディピティを考慮した CGM 小説推薦”, *人工知能学会合同研究会 2018 第 15 回データ指向構成マイニングとシミュレーション研究会*, 2018.
- [9] 高田叶子, 佐藤哲司, “文体の類似度を考慮したオンライン小説推薦手法の提案”, *DEIM Forum 2017 B5-2*, 2017.
- [10] 秦野智博, 阿部明典, “利用者の評価基準に合致した文章推薦システムの構築”, *2016 年度人工知能学会全国大会 (第 30 回)*, 2016.
- [11] Quoc V. Le, Tomas Mikolov, “Distributed Representations of Sentences and Documents”, *Proceedings of the 31st International Conference on Machine Learning(ICML 2014)*, pp.1188-1196, 2014.
- [12] Jey Han Lau, Timothy Baldwin, “An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation”, *Proceedings of the 1st Workshop on Representation Learning for NLP, Berlin, Germany*, pp.78-86, 2016.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, *arXiv preprint arXiv:1810.04805*, 2018.

物語内の人物と場所情報の時系列可視化による読書支援

Novel Reading Support by Time Series Visualization of Characters and Positions

MA Jiaxiu¹ * 西原陽子² 山西良典²
JIAXIU MA¹ Yoko Nishihara² Ryosuke Yamanishi²

¹ 立命館大学情報理工研究科

¹ Graduate School of Information Science and Engineering, Ritsumeikan University

² 立命館大学情報理工学部

² College of Information Science and Engineering, Ritsumeikan University

Abstract: This paper proposes a time-series visualization method of characters and positions in novel text. The proposed method will support users when they restart reading books. Once the already read part of a book is given, the proposed method extracts characters and positions every sentence and visualize them in time-series. As the intermediate report, we conducted experiments for a method of extraction of characters and positions. The averaged precision was from 0.88 to 0.98 and the averaged recall was from 0.79 to 0.97.

1 はじめに

インターネットの普及にともない、伝統的な紙の本の代わりに電子書籍が流通するようになった。電子書籍は1つの端末に複数の書籍を入れて持ち歩くことが可能である。複数の書籍を並行して読むこともしやすくなった。

長い時間を読書に取れない人は、1つの書籍を複数回に分けて読むこともある。複数の書籍を並行して読んでいると、1つの書籍の読書を中断してから再開するまでに長い時間がかかることもある。長い時間があいてしまうと、読書を再開するときに読んだ内容を忘れてしまうことがある。忘れたまま読み進めると、書籍の内容を把握しにくくなり、読書を十分に楽しむことができない。特に多くの人物が出てくるような物語であると、どの人物が誰とどこに居たのかが分かりにくくなる。既読部分の人物と場所の情報が可視化され、読書再開前に眺めることができれば、読んだ内容を思い出すことが容易になり、読書の支援ができると考えられる。

そこで本研究では、書籍の中の物語を対象とし、物語の既読部分から人物と場所情報を抽出し、時系列として可視化する手法を提案する。文から人物と場所を表す情報を抽出するだけでなく、人物が存在する場所を特定し、特定された情報を時系列として可視化する

ことに挑戦する。本稿では中間報告として、物語のテキスト内の人物と場所情報の抽出に対する評価結果を報告する。

2 関連研究

小説テキストを対象とした人物情報の抽出手法が提案されている [1, 2]。文献 [1] では、人名辞書「8万人西洋人名よみ方綴り方辞典」を利用し、辞書に載っている人名を小説テキストから抽出する。「性別」、「年齢」、「年代」、「職業」、「身体的特徴」、「性格」の六種類の人物情報を抽出し、人物リストを作成し、英米文学の推理小説を対象とし、人名の抽出を行う。本研究では人物リストにない人物情報を抽出するために、辞書ではなく、物語のテキストを機械学習することにより抽出を行う。

物語テキスト中の登場人物の関係を抽出する手法が提案されている [3]。人物関係を表す表現の辞書を作成し、表現に合致するパターンを構築し、パターンを用いて抽出を行う。本研究では人物間の関係ではなく、人物と場所の関係を扱う。誰がどこにいたのかの情報を可視化することにより、物語の進行を思い出しやすくする。

物語テキストから進行状況に応じて登場人物の存在状態と関係を推定する手法が提案されている [4]。存在状態を人物の生死に関わる「生存」、「死亡」、「死亡候補」の三種類とし、判定をする。本研究では登場した

*連絡先：立命館大学情報理工研究科
〒525-0058 滋賀県草津市野路東 1-1-1
nishihara@fc.ritsumeai.ac.jp

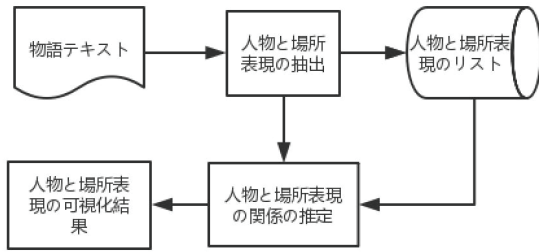


図 1: 提案手法の処理の流れ

人物の存在状態は扱わないが、既存研究と組み合わせることで、より詳細な時系列の可視化が可能になると考えられる。

3 提案手法

提案手法の概要を説明する。図 1 に提案手法の処理の流れを示す。入力は、物語のテキストである。ユーザの既読部分までを入力とする。入力された物語のテキストから、文ごとに人物と場所を表す文字列を抽出する。各文において、人物と人物がいる場所を紐づける。一文ごとに人物と場所の情報を可視化する。

本研究で理想とする可視化結果を図 2 に示す。物語「赤ずきんちゃん」の冒頭から 67 文目までの、人物と場所情報が正確に抽出された場合の可視化イメージ図である。

3.1 人物と場所表現の抽出

物語のテキストから人物と場所を表す表現を抽出する。人物と場所は固有表現の一種である。提案手法では固有表現抽出器の一つである Conditional Random Fields (CRF) [6] を用いる。CRF は条件付確率場のモデルであり、全体で最適な固有表現のためのタグ付けを行う手法である。CRF を用いた固有表現抽出器の作成手順を示す。

1. 物語のテキストを文ごとに分割する。
2. 各文に対して形態素解析を行い、単語と品詞情報を得る。
3. CRF で学習を行うために、単語に対し固有表現抽出のためのタグを付与する
4. CRF を用いて学習を行う。1つの単語に対し、自分自身と前後3つの単語、および品詞情報を学習し、固有表現抽出器を得る。

表 1: 提案手法で用いる BIO2 タグ。人物と場所表現用のタグを用いる。

表現タグ	説明
B-CHAR	人物表現文字列の始まる
I-CHAR	人物表現文字列が続いている
B-POS	場所表現文字列の始まる
I-POS	場所表現文字列が続いている
O	人物と場所表現以外の文字

表 2: 使用した物語のテキスト

番号	タイトル
1	赤ずきんちゃん
2	浦島太郎
3	桃太郎
4	猿蟹合戦
5	良夜

得られた固有表現抽出器を用い、物語のテキストから人物と場所表現を抽出する。物語のテキストは青空文庫などで公開されている物語のテキストを利用する。文末の句読点(、や。)、発話終了の鍵括弧など記号があれば文末と判定し、物語テキストの文への分割を行う。形態素解析器は MeCab、辞書は NEologd を用いる。

CRF での学習のためのタグは BIO2 形式 [5] を用いる。BIO2 では固有表現の先頭の形態素に B タグを付け、固有表現内の先頭以外の形態素に I タグを付与する。固有表現以外には O タグを付与する。本研究では人物と場所表現をそれぞれ区別して抽出するため、表 1 に示すタグを用いる。

CRF を用いて学習を行う際は、図 3 に示すように、 i 番目の形態素を中心として、前後3つの形態素をあわせた合計7つの形態素に対し、入力単語、品詞細分類、表現タグを用いる。

4 人物と場所表現抽出の評価実験

人物と場所表現を抽出する部分について適合率と再現率を用いて評価する実験を行なった。

本実験で用いた物語のテキストのタイトルを表 2 に示す。これらのタイトルのうち 1 から 4 は比較的日本人に読まれることが多いものとして選択した。5 のタイトルは、日本人には比較的知らない人が多いと思われるが、比較のために青空文庫の著者名順の先頭のものを選択した。

BIO2 のタグ付けは第一著者が行なった。タグづけされたデータを物語テキストごとに学習し、適合率と再現率を算出した。データは全て学習に用い、学習の際

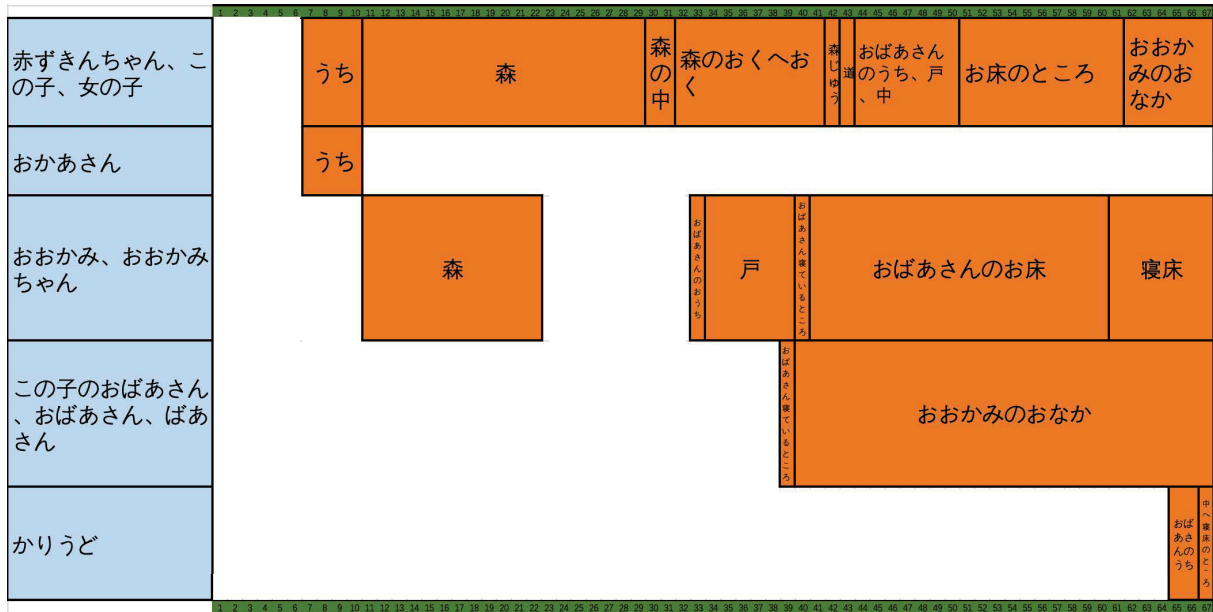


図 2: 物語「赤ずきんちゃん」の提案手法での可視化イメージ図。横軸に文の番号（緑色の長方形），縦軸に人物（水色の長方形），人物の登退場状態（オレンジ色の長方形），内部に人物が居る場所（オレンジ色の長方形中の黒字）を示す。

位置	入力単語	品詞細分類	表現タグ
i-3	これ	名詞-代名詞 一般	0
i-2	を	助詞-格助詞 一般	0
i-1	赤ずきん	名詞-固有名詞 一般	B-CHAR
i	ちゃん	名詞-接尾-人名	I-CHAR
i+1	,	記号-読点	0
i+2	ここ	名詞-代名詞 一般	0
i+3	に	助詞-格助詞 一般	0

図 3: CRF で学習する際の学習情報の例

の適合率と再現率を算出した。適合率と再現率は $B-CHAR$, $I-CHAR$, $B-POS$, $I-POS$ ごとに算出した。

4.1 実験結果と考察

算出された適合率と再現率を表 3 に示す。適合率の平均は 0.88 から 0.98 になった。再現率の平均は 0.79 から 0.97 になった。

CHAR タグ, POS タグのいずれにおいても, B タグの方が I タグよりも抽出における適合率と再現率が低かった。B タグは固有表現の冒頭の形態素に付与されるが, 今回の固有表現抽出では人物表現, 場所表現の冒頭の単語や品詞細分類が多様であったために, 適合率, 再現率がともに低くなったと考えられる。失敗例を表 4 に示す。失敗した例を表 4 に全て書いてもらえますか。物語ごとに分けてこれに対し I タグは B タグ

が出現した場合に後続するものであり, 条件が限定されやすい。このため I タグの適合率と再現率が B タグよりも高くなったと考えられる。

実験に用いた物語のテキストのうち 1 から 4 は適合率と再現率は高かったが, 5 のテキストは適合率と再現率が低かった。1 から 4 のテキストは子供向けの物語であった。5 のテキストは「良夜」という日本の物語で, 対象とする読者の年齢層は低くない。「良夜」の登場人物は「伯父」, 「父」が多く, これらの抽出に失敗することが多かった。人物名が具体的な名称（「赤ずきんちゃん」「浦島太郎」「桃太郎」など）ではないことがあり, 人物名と認識がされにくかったと考えられる。

5 おわりに

本研究では, 既読部分から人物と場所表現を抽出し時系列で可視化することにより, 読書を支援するための手法を提案した。本稿では手法の一部である, 人物と場所表現の抽出部分について評価実験を行い, 適合率と再現率により評価を行った。実験の結果, 適合率の平均は 0.88 から 0.98 になった。再現率の平均は 0.79 から 0.97 になった。子供向けの物語の方が人物, 場所表現の抽出が容易であることがわかった。今後は人物と場所の関係性を推定すること, 人物と場所の時系列可視化を行うことが課題である。

表 3: 人物と場所表現の抽出の適合率と再現率

物語	B-CHAR		I-CHAR		B-POS		I-POS	
	適合率	再現率	適合率	再現率	適合率	再現率	適合率	再現率
1. 赤ずきんちゃん	0.94	0.90	1.00	0.96	1.00	0.90	0.99	0.97
2. 浦島太郎	0.97	0.98	0.94	0.91	0.98	0.79	0.96	0.97
3. 桃太郎	0.96	0.97	0.95	1.00	1.00	0.89	1.00	1.00
4. 猿蟹合戦	0.98	0.98	1.00	1.00	1.00	0.87	0.97	0.97
5. 良夜	0.55	0.22	1.00	0.68	0.78	0.51	0.78	0.94
平均	0.88	0.81	0.98	0.91	0.95	0.79	0.94	0.97

表 4: 本手法の抽出失敗の結果例 (B タグと I タグそれぞれの失敗例を示し, 形態素解析の結果は記号" | "に分けた.)

物語	人物・場所	人手で付けたタグ	手法で付けたタグ
1. 赤ずきんちゃん	おばあさん	B-CHAR	O
	おばあさん の 着物	O O O	B-CHAR O O
	森 じゅう かけまわっ	B-POS I-POS O	B-POS I-POS I-POS
2. 浦島太郎	浜 べ	B-POS I-POS	O O
	海	B-POS	O
	かめの子	B-CHAR	O
3. 桃太郎	川	B-POS	O
	帰り	O	B-CHAR
	陸	B-POS	O
4. 猿蟹合戦	山道	B-POS	O
	かに	B-CHAR	O
	山 へ	B-POS O	B-POS I-POS
5. 良夜	父	B-CHAR	O
	新潟 県 下	B-POS I-POS O	B-POS I-POS I-POS
	伯父	B-CHAR	O
	猿	O	B-CHAR
	母	B-CHAR	O
	東京	B-POS	O

参考文献

- [1] 馬場 こづえ, 藤井 敦: 小説テキストを対象とした人物情報の抽出と体系化, 言語処理学会第 13 回年次報告, Vol. 13, pp. 574-577 (2007)
- [2] 米田 崇明, 崎 隆宏, 堀内 靖雄, 黒岩真吾: 述語情報を利用した小説の登場人物の抽出物語テキストを対象とした登場人物の関係抽出, 言語処理学会 第 18 回年次大会 発表論文集, Vol. 18, pp. 855-858 (2012)
- [3] 西原 弘真, 白井 清昭: 物語テキストを対象とした登場人物の関係抽出, 言語処理学会第 21 回年次大会発表論文集, Vol. 21, pp. 628-631(2015)
- [4] 縣 啓治, 伊藤 雄一, 高嶋 和毅, 北村 喜文, 岸野 文郎: 物語 テキストから進行状況に応じて登場人物の存在状態と関係を推定する手法, 第 18 回インタラクティブシステムとソフトウェアに関するワークショップ ,(2010)
- [5] 中野 桂吾, 平井 有三: 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, Vol. 45, pp. 934-941(2004)
- [6] 坪井 祐太, 鹿島 久嗣, 工藤 拓: 言語処理における識別モデルの発展 HMM から CRF まで, 日本 IBM 株式会社東京基礎研究所 グーグル株式会社,

言語学の語用論や配慮表現の先端技術への適用に関する一 考察

～敬語や待遇表現などの配慮表現は本当に必要か～

A Study on the Application of Linguistic Pragmatics and Consciousness Expressions to Advanced Technologies

太田 博三¹

Hiromitsu Ota¹

¹放送大学 教養学部 1

¹The Open University of Japan 1

Abstract: For dialogue systems, machine translation, and human agent interaction, it is illustrated and discussed whether consideration expressions such as honorific expressions and treatment expressions are really necessary. Artificial intelligence in recent years often refers to deep learning. However, knowledge-based artificial intelligence no longer appears in the table. Here, in this paper, the pragmatics and consideration expressions of linguistics are arranged, and how much is reflected and realized by advanced technology and how far it can be implemented is considered.

1. はじめに

国内外でのディープラーニングの進展は、スポーツ分野でのデータ解析への適用やコスモロジーなどの宇宙分野や天文学への専門分野への応用が試みられている段階にあり、応用段階にあると思われる。

一方、自然言語処理における機械翻訳の分野では、語用論の中の一つのポライトネスが必要との認識の下で、機械学習時の学習データ（教師データ）の準備の手間暇がネックとなっている。途中から、新しいタイプのデータを反映させるため、制御の効かないディープラーニングの分野では、小規模のテストが望ましいと考えられる。しかし、ポライトネスや敬語を中心とした待遇表現、または配慮表現の分野では、文化庁の国語審議会等を見る限り、ここ数十年での進展はあまり見受けられない。よって、技術面への適用も、例えば、ポライトネスを機械翻訳に反映させるべきなどの段階で停滞している。

本稿の位置付けは、ポライトネスや待遇表現、配慮表現を技術面に適用することが目的である。厳密な議論はその次の段階にある。

ポライトネスや待遇表現、または配慮表現に関

するコーパスを収集し、ユーザの運用を想定した、有効な分析手法を検討したものである。

2. 語用論の登場と技術面でのブレークスルー

2.1 先行研究

語用論では、ポライトネスが日本に紹介され、その位置づけや厳密な定義や捉え方において、様々な議論がなされてきた。結果的に次の2つに分かれて収束された。1つはグライスの発話論やゴフマンのフェイスの概念、そしてブラウン&レヴィンソン（以下、B&Lと記す）のポライトネス理論に乗っ取った捉え方であり、もう一つは、日本語の敬語と主体としたものである。ポライトネス理論や待遇表現、配慮表現は重複するものであるとの認識は否定的ではない段階でまともになっているように見受けられる。本研究では、生田(1997)の用いた配慮表現を主体的に用いるが、これに限らない。

2. 2 語用論の技術面への適用可能性先行研究

ポライトネスや待遇表現、配慮表現の捉え方が多岐に議論された中で、機械翻訳や対話自動生成への技術面への適用に限っては、アプリケーションや Web サービスでの提供になり、ユーザーからの評価に左右されるため、比較的、自由度が広いと考えられる。

これまででも、学問的な厳密性でとん挫した場合は、技術面への応用で、一気にブレークスルーすることがある。音声認識や自然言語処理など多岐にわたる分野がマツコロイドなどの対話ロボットに適用され、ユーザの評価が高まるのが正解となる。

2. 3 コーパスとコーパスの技術面への適用

国立国語研究所では、国家プロジェクトとして、『梵天』を除き、中立性を担保しつつ、大規模なコーパスが開発されている。一方、コーパスを機械翻訳や対話応答の自動生成に活かすには、正解としての教師データが必要不可欠であるため、csv 出力などデータとしての保存方法を統一し、ラベリングすることが簡易にできることが求められている。最終的には、ユーザーの評価、つまり、他人の使い勝手に帰属するため、機械学習にも適用可能なデータのあり方も分析手法への適用から有用なアウトプットも考慮されることになる。

3. ビジネス場面での配慮表現のストラテジーの分析とその妥当性について

コールセンターなどのビジネス面での会話につながるクレーム処理を取り上げる。服部(2009)は図 3-1 のように、大きく 3 つの段階に分けて、発話ストラテジーを捉えている。まずは、このように大雑把に分類し、次に、山岡(2017)の 2 種類の分類を追加することで、ラベルとして、教師データに近づけることが考えられる(図 3-2-1 と図 3-2-2 を参照のこと)。

一方、発話ストラテジーの捉え方が適切かについては、ベイズ統計学のベイズ更新を用いて、判

断する。4 節で、その詳細を検討する。

3. 1 先行研究

以下に配慮表現や待遇表現を含めた、ポライトネスの位置づけを示す。大きさは下記のような順のイメージである。

配慮表現 > ポライトネス > 待遇表現・敬語 > 語用論/ 敬意表現



図 3.1 ポライトネスに関する各定義の位置づけ

3. 1. 配慮表現とポライトネス

3. 1. 1. 1 塩田・山岡の定義～アンケート調査の視点から～

塩田(2012)では、配慮表現の定義を山岡(2010)の「対人的コミュニケーションにおいて、相手との対人関係をなるべく良好に保つことに配慮して用いられる言語表現」としている。この上で、配慮表現を定型的表現と自動詞表現・他動詞表現として捉え、アンケート調査を行い、クロス集計による男女差や世代差や地域差で認識の違いや傾向を述べている基本研究である。例えば、前者の定型的表現では、「つまらないものですが」はおかしな表現と捉えていたり、「おかげさまで」の支持は、女性(当時)の 50 歳に多く、男性は 49 歳以下に少ないなどである。

3. 1. 1. 2 松井の定義～関係性理論の視点から～

松井(2001)は、記号化されたポライトネスと紹介しており、ビジネスでは、一般に、クッション言葉としても周知されている。一方、後者では、

「皿が割れてしまいました (自動詞的表現)」と
 「皿を割ってしまいました (他動詞的表現)」と
 で、責任の主体がどこにあるかを示すものである。

3.1.1.3 松井の定義～原論的の視点から～

生田(1997)では、ポライトネス理論の紹介の中で、「ポライトネスは、当事者同士の互いの面子の保持、人間関係の維持を慮って円滑なコミュニケーションを図ろうとする社会言語行動を指す」として、この意味では、ことばのポライトネスは「配慮表現」、言語的「配慮行動」などと呼ぶ方が適切かもしれないと述べている。

3.1.1.4 山岡の定義～機能及び形式分類のコーパス的視点から～

山岡(2004)では、さらに、敬語もしくは敬意表現とポライトネスとの関係性を述べている。例えば、親疎関係による敬語の選択については、馴れ馴れしい表現を避けることで、相手との距離を保とうとする消極的ポライトネスの一種と見ることは可能。いわゆる上限関係の敬語の選択についても、目下の方より強い配慮が求められる日本語文化における独特な消極的ポライトネスと考えられる」と述べている。さらに、山岡(2017)では、ポライトネスの原理と配慮表現の機能分類との関係について明示している。しかし、すべてが対応している訳ではなく程度の問題である。部分的でその時の状況において、必ずしも適用されるものではない。一連のポライトネスと敬語または敬意表現とを同じものと見なす議論は、池上・河上(1987)のポライトネス理論を「丁寧さの原理」と訳したことに端を発し、井出(2011)が文化庁の国語審議会で敬語にポライトネス理論を応用したことに対する賛成反対の議論が、むしろ、ポライトネス理論に触れることを回避するような風潮にもなってしまった。

このように、海外で理論を学び、日本国内に輸入する際の語訳や誤解は、他の分野にもたくさん見受けられることであるが、時代の移行とともに風化し、整理されてゆくものでもあるように思われる。

3.1.1.5 金澤の定義～書き言葉における評価の視点から～

金澤編(2014)では、書き言葉コーパスとして用

いられている。12個のタスクの評価軸に、読み手配慮があり、「不快感を与えない表現」であるかなどの、読み手を配慮した表現か、また「敬語や終助詞の使用」が親しい間柄ならくだけた表現であるべきなどとなっている。ここでは、生田(1997)の例文(表3)と比較すると、ポライトネス理論はほとんど意識されていないように思われる。日本語教育の分野で、ポライトネスを考慮することは少なく、むしろ、混乱のもとになりかねない。日常生活での会話では、ポライトネス理論というより、語用論が多く存在し、松井(2001)では、命題を表意と推意に区分し、の発話の言語情報はとそこで意図して伝達する命題は変わらないとしている。

表 3.1.1.4-1 機能分類

I 利益表現
① 自利大 おかげさまで、ぜひ、～れば幸いです
② 他利小 つまらないものですが、何もありませんが、ご笑納ください
II 負担表現
① 他負大 お忙しいところ、わざわざ、お足元の悪いところ、お手数ですが、ご面倒ですが、可能なら、よろしかったら
② 自負小 ぜんぜん、大丈夫、喜んで、ついでに
III 緩和表現
① 侵害抑制 かもしれない、どちらかと言えば、言えなくもない
② 不一致回避 かもしれない、のほう、的には、とか
IV 賞賛表現
さすが、すごい、お見事、恰幅がいい、健康的、恐れ多くも
V 謙遜表現
① 自賛抑制 まだまだ、そこそこ、一応、自慢じゃないけど
② 自己非難 借越ながら、若輩者、不束者、出来の悪い、高い所から
VI 賛同表現
なるほど、たしかに、まったく、もの、ごもつとも
VII 共感表現
大変ですね、よかったですね、おつかれさま、それな、ですよー

表 3.1.1.4-2 形式分類

形式分類
① 副詞 ちよっと、ぜひ、ただ、まったく、たしかに、そこそこ、ぜんぜん、なるほど、そろそろ、なんか
② 副詞句 おかげさまで、悪いけど、すみませんが、恐縮ですが、失礼ですが、よろしかったら、ご多忙のところ、つまらないものですが、借越ながら、どちらかという
③ 形容詞・形容詞句 大丈夫、さすがですね、お見事、大変ですね、よかったですね
④ 接尾語・補助動詞 のほう、的な、的には、とか、なんか、～っぽい、～てくれる
⑤ 文末表現 ～でしょう、～かもしれない、～と言えなくもない、～のように思う、～ので(言いさし)、～ていただけるとありがたい、～てくれてもいいのに、～させていただきます
⑥ 慣用語 ご笑納ください、ご高名は伺っております、おつかれさまです

4. ポライトネスの例文による考察

本項では、生田(1997)と松井(2001)の例文を取り上げ、各々、主にポライトネスと敬語との違いを明示し、これらに対話生成に向けて背景や状況を考察する。

4. 1 生田(1997)の例文

状況は i から ii である。

- (i)「B のペンが何本か目の前の机の上に置いてある」
(ii)「相手 B が、既にペンを筆入れ、さらに鞆にしまっており、席を立とうとしている」

A1: 急いている?

B2: 別に。

A3: 悪いけど、ペン貸してくれる?

B4: いいよ(承諾の場合) / B: バイトに遅れそうなんだ。じゃあね(拒否の場合)。

A5: ごめん。せっかくしまったのに。

以下の状況のもとでの会話である。注目すべきは、この例では、敬語や丁寧な表現は一切使われていないことである。

ここで、A3 に、敬語を用いた表現を入れると不自然になる。

「A: あのを、恐れ入りますが、ちょっとペンをお借りできますでしょうか。」

これらから、ポライトネス(語用論)と敬語は異なるものだと言える。よって、ポライトネスの表現をコーパスに取り入れられているか吟味する必要がある。

4. 2 松井(2001)の例文

1 対の対話でしかないが、これを推論するのに

は、単なる文の省略を補完するだけでは、十分である。背景や親子関係も考慮する必要がある。

背景は(カ)から(コ)にかけて記す。

(カ) 美佐の母親が今日自宅のベランダにふとんを干していた。

(キ) 美佐は美佐の母親が今日自宅のベランダにふとんを干していたと言った。

(ク) ふとんを干すのは、宿泊する客が来る前の準備である。

(ケ) 美佐の母親は真理のためにふとんを干していた。

(コ) 美佐の母親は真理が明日泊まりに来るのを知っている。

真理「お母さん、私が明日泊まりにゆくの知ってるの?」

美佐「ふとんを干してたわ」

また、ここでは、表意と推意とに分けて捉えている。発話の言語情報はとそこで意図して伝達する命題は変わらないとしている。敬語や丁寧な表現は一切使われていないことも確認できる。

4. ベイズ論的アプローチによる会話ストラテジーの方法論の検討

本節では、図 4.5.1 のように、応答詞(「はい」や「はいはいはい」と笑い(<笑>)の出現数を1回目から5回目まで、別々にカウントし、「笑い」が出て「はい」が出ている場合のポライトネス理論の水準を次の3つに定めて、モデル化したものである。

1) ポジティブ

2) ゼロ

3) ネガティブ

次に、尤度を算出する

その次に、事前確率を設定する。ここでは、理由不十分の原理から事前確率を等しく設定した。

さらに、カウントしたデータを入力し、事後確率を計算した。

最後に、データを変えてみる。これらをグラフ

に図示した。

配慮表現に該当する語句や品詞の組み合わせの出現確率を計算し、語句とストラテジーの流れをモデル化し、その傾向性を検討したものである。手順は以下のとおりである。

まず、水準を次の3つに定めて、モデル化する。

- 1) ポジティブ、2) ニュートラル(ゼロ)、3) ネガティブ

次に、尤度を算出し事前確率を設定する。ここでは、理由不十分の原理から事前確率を等しく設定する。

「ちょっと」とく申し訳ございません>をベースで推定				
(1)「ちょっと」とく申し訳ございません>のモデル化				
		ポジティブ(H ₁)	ゼロ(H ₂)	ネガティブ(H ₃)
憎因子の数		3	3	3
愛因子の数		4	2	3
(2) 尤度の算出				
データ(D)		P(D H ₁)	P(D H ₂)	P(D H ₃)
悪印象(憎因子選択)		0.429	0.600	0.500
良印象(愛因子選択)		0.571	0.400	0.500
(3) 事前確率の設定				
		ポジティブ(H ₁)	ゼロ(H ₂)	ネガティブ(H ₃)
最初の事前確率		0.3	0.4	0.3
(4) データ入力と事後確率の算出				
データ回数	データ(D)	ポジティブ(H ₁)	ゼロ(H ₂)	ネガティブ(H ₃)
		0.300	0.400	0.300
1回	良	0.414	0.276	0.310
2回	悪	0.414	0.276	0.310
3回	良	0.527	0.176	0.297
4回	悪	0.527	0.176	0.297
5回	良	0.630	0.105	0.266
6回				
7回				
8回				
9回				
10回				

図 4.5.1 尤度による表現の出現順序の計量化案

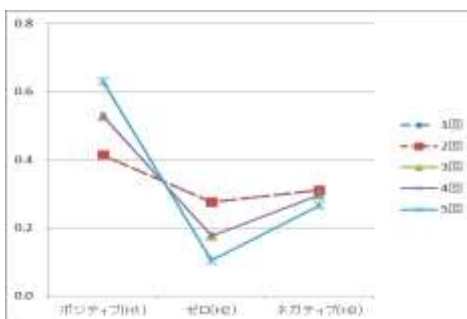


図 4.5.2 尤度のポジネガ変化

4. 2 クレーム対応のストラテジーの事例考察

主に、コールセンター等でのクレーム対応は、次の3つから構成される。

- 1) お客様のご状況を伺う「前段」、
- 2) 相互状況の確認とお詫び
- 3) 先々に向けたクロージングの挨拶

表 4.2 クレーム対応のスキプトの一例

01 R:	えと私がお持ちできるかちょっと他の担当者(あの:になるか)が分からないんですが()あの:まず商品の在庫のほうだけ私のほうから御連絡入れさせていただきますので: =
02 C:	=はい
03 R:	え()え()であの:また別にちょっと担当者のほうからはC様へお電話入れさせていただくようにはしますので=大変申し訳ございません。ご迷惑おかけいたしました=
04 C:	=いえ、あ()あの:()できるだけ早くお願いします=
05 R:	=じゃああの:在庫のほうの確認だけして先にお電話させていただきます。よろしいです[か、私JB01と申します]ので
06 C:	[あ()]はい、はいR様で
07 R:	何かありましたらまた御連絡くださいませ
08 C:	あ()はい、よろしくお願ひいたします。
09 R:	[申し訳ございません、よろしくお願ひいたします。はい[失礼いたします]
10 C:	[はい=では失礼いたします。]

※ 記号Rは受け手(JB)、Cはかけ手

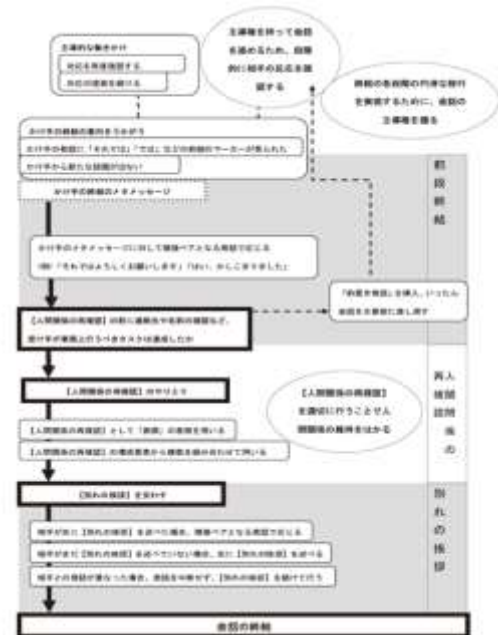


図 1 ビジネス場面におけるクレーム電話会話経路

図 4.2 クレーム対応のストラテジー

5. まとめ・今後の展望

本稿では、ポライトネスの表現を抽出し、コーパスを作成する際に、配慮表現や敬語を含めるか否かで、壁にぶつかるため、コーパスの範囲を定めるために、一度、整理し、各々の違いを明確にした上で、推論による実装に近づけた。Lisp や Prolog での推論を試みたが、十分な成果は得られなかった。文の省略の補完では不十分であり、文と文とのつながりがポイントになるからであることが、改めて、わかった。

これらの試行錯誤の末、オントロジーの推論チャレンジのように、背景や人間関係の近さなどを表現する必要がある。これは、別途、小説のように、会話部分と描写とが区分されつつも、同じナレッジスクリプトに置かれるのが好ましいと考えられる。

コンピューターを用いる醍醐味は、人間の頭脳では、なかなかわかりにくい事項を指し示すことであるため、オントロジーをベースにした意味ネットワークを構成し、推論でポライトネスの文が出せるか、検討中であり、成果は次回以降に示したい。

参考文献

- 1) 生田少子(1997). 「ポライトネスの理論」大修館書店, 『月刊言語』, pp.66-71.
- 2) 彭飛(2004). 『日本語の「配慮表現」に関する研究—中国語との比較研究における諸問題』 和泉書院
- 3) 山岡政紀(2017). 「日本語配慮表現の分類と語彙リストについて」『日本語コミュニケーション研究論集』第7号, 日本語コミュニケーション研究会, 3-11
- 4) 山岡政紀(2004). 「日本語における配慮表現研究の現状」 日本語日本文学, 創価大学紀要論文
- 5) 服部明子(2009)「電話会話における日本人ビジネス関係者のクレームへの応対」『言葉と文化』(10), 77-93, 2009-03
- 6) 松井智子(2001), 「関連性理論から見たポラ

- イトネス」 30 卷 No.3 特集 月刊言語
- 6) 久野暲 (1978)『談話の文法』大修館書店
- 7) 金澤裕之編(2014), 「日本語教育のためのタスク別書き言葉コーパス」ひつじ書房
- 8) 堂下・白井・溝口・新美・田中(1998)『音声による人間と機械の対話』「対話過程の基本的特徴と対話における省略の処理(第2章)」オーム社
- 9) ペネロピ・ブラウン・スティーヴン・C・レヴィンソン(1987)「ポライトネス 言語使用における、ある普遍現象」 “Politeness: Some Universals in Language Usage”
- 10) 瀧田・西島編著 (2019). 「機械翻訳と未来社会—言葉の壁はなくなるのか」社会評論社, pp.53-104.
- 11) 松原望(2008). 「入門ベイズ統計—意思決定の理論と発展」東京書籍

関連 URL

- 1) 『国語研日本語ウェブコーパス』検索系『梵天』 <http://bonten.ninjal.ac.jp/>
- 2) 文化庁の国語審議会：現代社会における敬意表現(2000年12月8日) http://www.mext.go.jp/b_menu/shingi/old_bunka/kokugo_index/toushin/1325322.htm
- 3) ナレッジグラフ推論チャレンジ <https://challenge.knowledge-graph.jp/2019/>

補足

久野(1978)は、英語の省略現象をもとに、省略と談話法規則とに言及し、省略の順序について述べている。より新しい(重要な)インフォメーションが残され、より古いものが省略されるとしている。また談話法規則として、省略の根本原理に、省略されるべき要素は、言語的あるいは非言語的文脈から復元可能でなければならないとしている。堂下・白井・溝口・新美・田中(1998)は、音声対話システムの実装に向けて、対話における省略を、次の2つに分けている。

- 1) 対話当事者に関する省略(「私は」や「あなたに」など),
 - 2) 「共有知識に基づく省略(文脈中には言及されていないが、対話の当事者が共有している知識に基づいて、省略が行われるもの)」
- この分類の上で、省略されている名詞句、修飾詞句、動詞句を省略詞区とし、この頻度を計算し、補

完するものである。その省略文の前の文を 0-6 の範囲として、頻度を測定し、67.9%の割合で直前の文に省略詞があると結論付けている。伝 (1997) は、話し言葉の特徴を、言い淀みと言い直し、省略とし、音声対話コーパスでの統一モデルを提案し、言い淀みや言い直しなどの不適合性を適切に扱う手法について述べている。

SNS上の悪口を含む投稿に対する取り下げを促す フィードバック文の自動生成方法の検討

A Study on Automatic Feedback Message Generation for Asking to Withdraw SNS Message Including Toxic Expression on SNS

藤堂悠杜^{1*} 西原陽子¹ 山西良典¹
TODO, Yuto¹ NISHIHARA, Yoko¹ YAMANISHI, Ryosuke¹

¹ 立命館大学 情報理工学部

¹College of Information Science and Engineering, Ristumeikan University

Abstract: In this paper, we conducted experiments to obtain the most effective feedback sentence to ask for withdrawing SNS messages including toxic expressions. In our study, Cognitive Response Model and use self-talk feedback to SNS users. We prepared four types of self-talk feedback sentence and experimented the four with questionnaires on the Web. We set four types of condition in obtaining a feedback sentence: whether or not a SNS message includes explicit toxic expression, and whether or not the feedback sentence points out the toxic expression. Experimental results showed that the most effective feedback was a sentence that presented a sender to a receptor of toxic expression. We also found that the self-talk feedback was more effective when a SNS message included an explicit toxic expression and the feedback pointed out the toxic expression.

1 はじめに

近年のインターネットの普及やスマートフォンの発達などの影響より、小・中学性からインターネットに触れる機会が増加している。総務省の調べ [1] によると、2011 年では 13 歳から 19 歳におけるスマートフォン所有率は 14.6 パーセントだったのに対し、2016 年には 81.4 パーセントにまで上昇している。こうしたネット社会の普及によって、利便性が増す一方で、インターネットの危険な側面も広がってしまっている。その一例がインターネットでのいじめであり、新聞記事などでもその被害が報告されている。

インターネット上でのいじめについての研究、それを防ぐ研究は数多く行われている [2, 3, 4, 5]。インターネット上でのいじめにおいては、被害者にとって有害な情報を投稿することにより行われることがある [6]。インターネット上では匿名で発言できることや、相手の顔が見えないことから、攻撃的な投稿が生まれやすくなってしまふ、いわゆるフレーミング現象が起こりやすくなっている [7]。有害表現を含む投稿をフィルタリングする技術が研究されているが [8]、有害表現を含む投稿がなされてしまうことは防ぐことは難しい。2019 年 6 月には Twitter において、有害な表現を含むリプライ

が送られると、リプライを受けたユーザとそのフォロワーを含む全員からリプライが隠される機能が実装された [9]。これにより、有害表現を含むリプライは投稿者以外からは見ることができない状態になった。さらに、Instagram でも攻撃的な内容を含む投稿を検知すれば、“Are you sure you want to post this?”と、ユーザを諭す機能を実装した [10]。有害表現を含む投稿が、被害者やその周りの人の目に触れないようにする取り組みが進められている。

ただ一方で、有害表現を含む投稿を一方的に規制してしまうと、より一層攻撃的な投稿をするユーザの気持ちを刺激してしまい、周りから見えないからこそ言いたい放題な投稿をしてしまう可能性も考えられる。それではインターネットを利用する情報モラルの観点で考えると根本的解決に至っていない。ネットいじめに対して情報モラル教育の拡充の必要性も示されており [11]、ユーザ自身の判断で、投稿を取り下げさせることも必要と考えられる。さらに、Instagram などの手法では、示されるフィードバック文が常に同じものである。投稿内容に応じてフィードバック文を生成し、示すことにより、有害な表現を含む投稿の取り下げに対して効果が高まると考えられる。

本研究では SNS 上の有害表現を含む投稿に対し取り下げを促す方法を提案することを目的とする。このために、本稿においては、有害表現のうち他者への悪口

*連絡先：立命館大学情報理工学部
滋賀県草津市野路東 1-1-1
E-mail: nisihara@fc.ritsumeai.ac.jp

を扱い、悪口を含む投稿の取り下げを促すことに有効なフィードバック文について実験により明らかにする。フィードバック文は単に投稿を禁止するものではなく、Greenwaldの「説得の認知反応プロセス」[12]を参考にし、投稿者にセルフトークを起こさせるものとする。本研究が想定する投稿者は、他者に対して攻撃的な投稿（脅迫のような違法な情報[13]）をしようとしている人ではなく、他者からの指摘を受けると行動を省みる人である。これにより、不用意に他者を傷つける発言をしてしまう投稿者自身も保護したい。

2 関連研究

はじめに、態度変容についての関連研究を紹介し、本研究の位置付けを行う。Greenwaldは「説得の認知反応プロセス」において、人はメッセージを受け取ることによって態度変容を起こすのではなく、メッセージを受け取った後に行うセルフトークにより態度変容を起こすとしている[12]。セルフトークとは自己会話であり、自分自身に対して話しかけることである。既存のSNSサイトでは投稿の確認をする際に、セルフトーク形式のフィードバック文を提示するものは少ない。本研究では、セルフトーク形式のフィードバック文を採用し、最も効果の高い文面について明らかにする。

五十嵐らの研究においては、ドライビングシュミレータを用いて速度違反を抑止する効果的なメッセージについて検証が行われた[14]。速度違反や遵守した際にメッセージが提示される場合と、速度にかかわらず、常時メッセージが提示される2パターンの提示方法を用いられた。メッセージの文面としては禁止型と感謝型の2パターンが用いられた。検証した結果、禁止型および感謝型のいずれにおいても、速度抑止の効果に違いは見られないが、禁止型よりも感謝型のメッセージの方が、時間が経過しても効果が持続するということがわかった。禁止型だと心理的リアクタンス、すなわち制限されると反発したくなる気持ちが働いたことが影響したためと考えられる。清の研究[15]においても、医師が患者の行動を制する場合には、「～しないでください」と直接的な禁止表現ではなく、「～はやめておきましょうか」と共感性の高い禁止表現を用いることが多いことが確認されている。これらのことから、文面によっても、態度変容への影響が変わると考えられる。本研究ではセルフトーク形式の文面を、他者への自己投影型と未来予想を提示する型の2つに分け、その効果を検証する。

3 取り下げに有効な フィードバック文の調査実験

取り下げに有効なフィードバック文の調査を行うべく、被験者実験を行なった。実験では、被験者を募り、アンケートに回答することを依頼した。アンケート結果を分析し、取り下げに有効なフィードバック文を調査した。

3.1 実験目的

本実験では、以下の3点を明らかにすることを目的とする。

1. 悪口の具体的な指摘は取り下げに寄与するか？
2. 悪口の種類の取り下げに寄与するか？
3. セルフトークの種類は取り下げに寄与するか？

1つ目で明らかにしたいことは、SNSへの投稿文の中に悪口が含まれるときに、その悪口の単語またはフレーズを指摘することで取り下げ率が上がるかどうかである。情報システムにより提示されるメッセージは無視されがちであるが、悪口の単語を具体的に指摘することで、無視されることが減り、結果的に取り下げ率が上がる可能性がある。

2つ目で明らかにしたいことは、SNSへの投稿文の中に含まれる悪口が露骨な悪口の場合に、取り下げ率が上がるかどうかである。露骨ではない悪口（隠語など）の場合は意図して書いていると考えられるが、露骨な悪口は思いもよらず書いてしまうこともあり、指摘をすると書いてしまったことに気づけ、結果的に取り下げ率が上がる可能性がある。

3つ目で明らかにしたいことは、セルフトークを起こさせるメッセージの形式が、他者へ自己を投影する形式か、未来の利益を提示する形式かで、取り下げ率が上がるかどうかである。他者へ自己投影することにより、他者への共感が生まれ、悪口投稿を自分が受け取った場合を考える、つまりセルフトークを起こすことができ、結果的に取り下げ率が上がる可能性がある。一方で、未来の利益を提示する場合は、自己の利益を受け取った場合を考える。こちらもセルフトークを起こすことができ、結果的に取り下げ率が上がる可能性がある。他者への共感は生まれにくいいため、セルフトークは起こりにくいと考えられる。

以上の3つのことを明らかにすることを、本実験の目的とする。

表 1: セクションごとの条件設定

セクション	セルフトーク	露骨な悪口	悪口の指摘
1	○	○	○
2	○	○	×
3	○	×	○
4	○	×	×

3.2 実験手順

実験は次の手順で行なった。初めに実験者（第一著者）が被験者を募集する。被験者が SNS へ悪口を含む投稿を試みていると仮定し、被験者には悪口の投稿文と投稿文の取り下げを促すフィードバック文が示される。被験者は投稿文とフィードバック文を見て、投稿文を取り下げるかどうかを回答する。

投稿文の取り下げについての質問は、悪口の指摘の有無と、露骨な悪口か否かで条件を変え、4回のセクションに分けて行われた。表 1 に各回のセクションでの条件設定を示す。全てのセクションにセルフトーク形式のフィードバック文が提示される部分は共通しているが、悪口の指摘の有無と露骨な悪口か否かの条件がそれぞれ異なっていた。セクション 1 からセクション 4 まで、被験者はそれぞれ異なる順番で回答をした。つまり、ある人はセクション 1,2,3,4 とたどり、別の人は 1,3,4,2 と辿るなど、被験者ごとにたどるセクションの順番が異なっていた。

3.2.1 悪口の投稿文とフィードバック文の提示方法

SNS でグループチャットを行っているときに、被験者自身が他のメンバーの 1 人に向けて悪口の投稿を試みているという状況を仮定した。フィードバック文が提示された時に、投稿を取り下げるかどうかを回答してもらった。

グループチャットの人数は 3 名とした。グループチャットの投稿文は、チャットにおけるネットいじめが体験できる Web サイトにあるログを参考に作成した¹。図 1 に、実験で用いたグループチャットの例を示す。図は、Web 上でチャット画面を作成することができるサイトを用いて作成した。² 右側から出ている吹き出しが被験者の投稿文、ユーザ A が一緒に悪口を投稿しており、ユーザ B が悪口を言われている人となる。

悪口の投稿文は露骨な悪口を含むものと、隠語の悪口を含むものを用意した。露骨な悪口を含む投稿文は図 1 の最後に、「今まで黙ってたけどマジでお前馬鹿すぎてもう無理。死ねよ。」で、「馬鹿」「死ね」と露骨な悪口の単語を含んでいた。こちらの投稿文も同じサイト

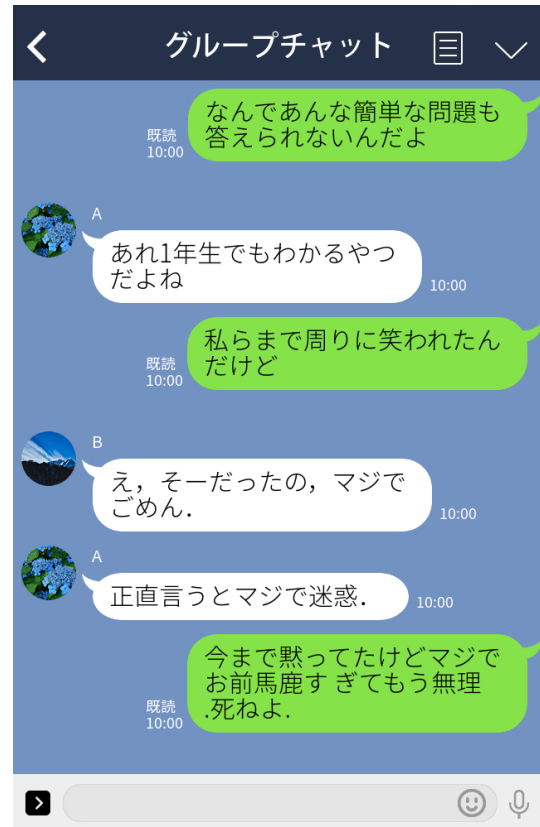


図 1: 実験で用いたグループチャットの例。

のログを参考に作成をした。隠語の悪口を含む投稿文は図 2 の最後に、「この集合写真の A さん、目が半開きで笑う（笑）」で、露骨な悪口の単語やフレーズは含まれないが、文全体として A さんへの悪口になっていた。

提示するフィードバック文は 1 つは他者への自己投影形式であり、もう 1 つは未来の利益を提示する形式である。これらはさらに 2 つずつ分け、合計 4 種類のフィードバック文を用意した。表 2 に実験で用いたフィードバック文を示す。

- 他者へ自己投影する形式：
 - 1 悪口の対象を投稿者に置換したフィードバック
 - 2 周りや相手への影響を示唆するフィードバック
- 未来の予想を提示する形式：
 - 3 投稿者の未来を予想するフィードバック
 - 4 グループチャットの未来を予想するフィードバック

フィードバック文を見て取り下げるかどうかについて、5段階の選択肢を用意し、選択肢の中から 1 つ選ぶことで回答をしてもらった。5段階の選択肢は、投

¹<http://linelog.jp/> (2019 年 10 月 31 日確認)

²<https://sp.mojimaru.com/> (2019 年 11 月 2 日確認)



図 2: 実験で用いたグループチャットの例.

稿を取り下げたいと「強く思う: +2」「少し思う: +1」「どちらとも言えない: 0」「あまり思わない: -1」「全く思わない: -2」とした.

3.2.2 被験者の情報

本実験の被験者は情報理工学部に所属する 20 代の男女 16 名で、内訳は男性が 11 名、女性が 5 名であった。共感的関心についての質問、およびセクション 1 から 4 までの質問に回答し終えた後、被験者から実験の感想をもらった。

表 2: 実験で用いたフィードバック文

分類	フィードバック文
1 (自己投影)	「もしその投稿をあなたが受けたら、あなたは不快に思いませんか？」
2 (自己投影)	「あなたの投稿は、トークルームの人や、相手を不快にさせていませんか？」
3 (未来提示)	「もし投稿を取り下げれば、周囲や相手を傷つけずに済みますよ？」
4 (未来提示)	「あなたの投稿の後、その先のトークはどう進んでいくと思いますか？」

4 実験結果と考察

表 3 に、アンケートの調整を行なった後の評価値の平均を取ったものを示す。各行、被験者 16 名分の評価値の平均となっている。評価値が高いほど、取り下げに有効であったことを示している。評価値が高かったものは、セクション 1 におけるフィードバック 1、つまり露骨な悪口が投稿文に含まれ、その悪口を具体的に指摘し、「もしその投稿をあなたが受けたら、あなたは不快に思いませんか？」とフィードバック文を提示した時であった。続いて評価値が高かったものは、セクション 2 におけるフィードバック 2、つまり露骨な悪口が投稿文に含まれるが、その悪口の指摘はせず、「あなたの投稿は、トークルームの人や、相手を不快にさせていませんか？」とフィードバック文を提示した時であった。フィードバック 2 はセクション 1 においても高い評価値を示した。フィードバック 1 と 2 は、セクション 3 を除き、正の評価値が得られており、取り下げに有効な可能性があることがわかった。

反対に評価値が低かったものは、セクション 4 におけるフィードバック 4、つまり露骨な悪口も、悪口の指摘もなく、「あなたの投稿の後、その先のトークはどう進んでいくと思いますか？」であった。フィードバック 4 はセクション 3 においても低い評価値を示した。続いて評価値が低かったものは、セクション 4 におけるフィードバック 3、つまり露骨な悪口ではないが、悪口を指摘し、「もし投稿を取り下げれば、周囲や相手を傷つけずに済みますよ？」であった。フィードバック 3 と 4 は全てのセクションにおいて負の評価値が得られており、取り下げに有効ではない可能性が高いことがわかった。

4.1 悪口の具体的な指摘に対する取り下げの効果

実験目的の 1 に対応し、悪口の具体的な指摘に対する取り下げの効果について検証する。セクション 1 と 2 は、露骨な悪口を含む投稿文であることは共通して、悪口の指摘の有無が異なっている。セクション 1 の評価値の平均は 0.187 であり、セクション 2 の評価値の平均は 0.047 であった。指摘をする方が評価値の平均は高くなった。また、セクション 3 とセクション 4 は、露骨ではない悪口を含む投稿文であることは共通して、悪口の指摘の有無が異なっている。セクション 3 の評価値の平均は -0.422 であり、セクション 4 の評価値の平均は -0.250 であった。指摘をしない方が評価値の平均は高くなった。このことから悪口に相当する単語を具体的に指摘することで、投稿の取り下げに効果があるが、露骨な悪口に限定される可能性が明

表 3: 各設問におけるアンケートの評価値

セクション	フィードバック	評価値
1	1	0.812
	2	0.375
	3	-0.437
	4	0
	平均	0.187
2	1	0.312
	2	0.500
	3	-0.375
	4	-0.250
	平均	0.047
3	1	-0.250
	2	-0.187
	3	-0.562
	4	-0.687
	平均	-0.422
4	1	0.250
	2	0.312
	3	-0.687
	4	-0.875
	平均	-0.250

らかになった。被験者からの感想でも、「有害ワードが具体的に示されるほうが、なにがいけなかったのかわかりやすいため。」や、「具体的な忠告によって、一旦メッセージを全て読み、考えることができる。ありきたりの文であると、すぐにその警告文を無視してしまう。」とあった。

4.2 悪口の種類に対する取り下げの効果

実験目的の2に対応し、悪口の種類に対する取り下げの効果について検証する。セクション1と3は悪口の指摘は共通しているが、悪口の種類が異なっている。セクション1の平均値は0.187であり、セクション3は-0.422であった。露骨な悪口の方が評価値が高くなった。セクション2と4は悪口の指摘をしない点で共通しているが、悪口の種類が異なっている。セクション2の平均値は0.047、セクション4の平均値は-0.250であった。こちらも露骨な悪口の方が評価値が高くなった。このことから、露骨な悪口の方がフィードバック文を提示することにより、取り下げられる可能性が高いことがわかった。自由回答からは、悪口かそうでないか微妙なものに対してフィードバックを行っても、ユーザ自身にそもそも悪い投稿を行っているという自覚がない場合には、「何が投稿してはいけなくて注意されているのかわからない」という意見や、「機械の誤認識も

表 4: セクション1で得られた評価値に対し、平均値の差の検定を行い得られた p 値。bonferroni 法による補正後値

	FB1	FB2	FB3
FB2	0.8990	-	-
FB3	0.0017	0.1629	-
FB4	0.1933	1.0000	0.1770

考えられるから効果が無い」という意見が見られた。

4.3 セルフトークの種類に対する取り下げの効果

実験目的の3に対応し、セルフトークの種類に対する取り下げの効果について検証する。ここまでの実験結果に対する考察において、露骨な悪口に対し、具体的な指摘をすることで、フィードバック文を示すと、取り下げに対し効果があることがわかった。そこで、本節ではこの条件に合致するセクション1で得られた評価値に対し、4種類のセルフトークの中で取り下げに効果があるものについて考察をする。

セクション1で得られた調整後の評価値に対し、フィードバック文のペアごとに平均値の差の検定を行った。多重比較となるため bonferroni 補正を行なった。検定の結果を表4に示す。有意差が見られたのは、フィードバック文1と3の間であった ($p=0.0017 < 0.05$)。この結果からフィードバック文3と比較すると、フィードバック文1は取り下げに対し、効果が高いことが明らかになった。フィードバック文1は自己投影型で、フィードバック文3は未来提示型であった。悪口を受ける対象を自己に置き換えることで、相手の気持ちに共感することができたため、取り下げに効果があったと考えられる。

5 おわりに

本稿では、SNS上における悪口を含む投稿に対し、取り下げを促すフィードバック文を自動生成するために、取り下げ効果の高いフィードバック文を実験を用いて明らかにした。フィードバック文は投稿を単に禁止する文面ではなく、Greenwaldの「説得の認知反応プロセス」で挙げられているセルフトークを起こさせる文面とした。効果の異なる文面を4パターン用意し、最も取り下げの効果が高いものを実験により明らかにした。実験の結果、他者への自己投影を起こす形式の文面の取り下げに対する効果が高いことがわかった。また、実

験において、投稿文に含まれる悪口の種類が取り下げに寄与するか、フィードバック文内で悪口の具体的な指摘が取り下げに寄与するかについても調べた。結果として、投稿文に露骨な悪口が含まれた時に、フィードバック文を提示すると取り下げ率が高くなることがわかった。さらに、フィードバック文で悪口を具体的に指摘することでも取り下げ率が高くなることがわかった。一方で、この結果は大学生から得られた結果であり、小中学生など、発達段階が異なる層に対しては異なる結果が出る可能性がある。今後、同様の実験を行い、発達段階の違いを考慮した上で、取り下げ効果の高いフィードバック文について明らかにする。

謝辞

本研究の一部は、科研費(17K13254)、安心ネットづくり促進協議会の助成を受けて行われました。記して謝意を申し上げます。

参考文献

- [1] 情報通信白書平成30年版。総務省, 2018.
- [2] 笹川喬介, 和泉順子. 誹謗中傷問題のインターネットによる影響に関する考察. 情報処理学会研究報告(グループウェアとネットワークサービス), 第2013-GN-89巻, pp. 1-6, 2013.
- [3] 藤桂, 吉田富二雄. ネットいじめ被害者における相談行動の抑制. 教育心理学研究, Vol. 62, No. 1, pp. 50-63, 2014.
- [4] 内海しよか. 中学生のネットいじめ, いじめられ体験—親の統制に対する子どもの認知, および関係性攻撃との関連—. 教育心理学研究, Vol. 58, No. 1, pp. 12-22, 2010.
- [5] Michele L Ybarra and Kimberly J Mitchell. Youth engaging in online harassment: associations with caregiver-child relationships, internet use, and personal characteristics. *Journal of adolescence*, Vol. 27, No. 3, pp. 319-336, 2004.
- [6] 田代光輝, 服部哲. 情報倫理 ネットの炎上予防と対策. 共立出版, 2013.
- [7] 加藤由樹, 加藤尚吾, 杉村和枝, 赤堀侃司. テキストコミュニケーションにおける受信者の感情面に及ぼす感情特性の影響: 電子メールを用いた実験による検討. 日本教育工学会論文誌, Vol. 31, No. 4, pp. 403-414, 2008.
- [8] Ryuichi Omi, Yoko Nishihara, and Ryosuke Yamanishi. Extraction of paraphrases using time series deep learning method. In *International MultiConference of Engineers and Computer Scientists 2019*, pp. 276-278, 2019.
- [9] *Giving you more control over your conversations*. <https://blog.twitter.com/enus/topics/product/2019/morecontrolofconversation.html>, 2019.
- [10] *Our Commitment to Lead the Fight Against Online Bullying*. <https://instagrampress.com/blog/2019/07/08/our-commitment-to-lead-the-fight-against-online-bullying/>, 2019.
- [11] 三島浩路, 本庄勝. 技術的観点からのネットいじめ対策. 電子情報通信学会 通信ソサイエティマガジン, Vol. 9, No. 2, pp. 102-109, 2015.
- [12] Anthony G. Greenwald. *Cognitive Learning, Cognitive Response to Persuasion, and Attitude Change*, chapter 6, pp. 147-169. Academic Press Inc., 1968.
- [13] 小向太郎. インターネット上の青少年犯罪被害対策の動向. 情報処理学会研究報告(グループウェアとネットワークサービス), 第81巻, pp. 1-6, 2011.
- [14] 五十嵐彩那, 白井伸之介. 速度違反抑制に効果的なメッセージと提示タイミング. 交通科学, Vol. 46, No. 1, pp. 13-20, 2015.
- [15] ルミ清. 禁止の場面における現実の言語表現: 医師と美術館員の場合. 世界の日本語教育: 日本語教育論集, Vol. 16, pp. 107-123, 2006.

階層的トピックモデルに基づく宿泊レビューの時間変化分析

Time-Variable Analysis of Accommodation Reviews Based on Hierarchical Topic Model

佐藤 裕次郎 山西 良典 西原 陽子 *
Yujiro Sato¹ Ryosuke Yamanishi² Yoko Nishihara²

¹ 立命館大学大学院情報理工学研究科

¹ Graduate School of Information Science and Engineering, Ritsumeikan University

² 立命館大学情報理工学部

² College of Information Science and Engineering, Ritsumeikan University

Abstract:

Accommodation review is a valuable resource for future users to know the voices from the users who have already stayed there. However, it is difficult for users to extract the information specific to each topic such as facilities, access, and breakfast. We consider that seasonal features of accommodation are especially important for comfortable and enjoyable staying. This paper proposes the hierarchical topic analysis with time variation to extract seasonal features for accommodations. The proposed method extracts seasonally important words and shows the similarity of topics that the important words belong to between seasons. In this paper, we discuss the effectiveness of the extracted features for references to choose accommodations.

1 はじめに

宿泊予約の主流がネットになったことで、インターネット上の情報をもとに宿泊施設を決定する能力が宿泊検討者に必要とされるようになった。宿泊施設の決定要因となる要素は様々であり、ユーザによって基準は異なるものの、Dickingerらの研究によると、友人からの推薦とオンラインレビューがネット予約において最も重要な影響を与える要因であると示されている [1]。その中でも宿泊レビューは幅広く研究されており、テキストマイニングによる分析結果を意思決定に役立てる研究が進んでいる。Vermeulenらの研究によると、宿泊レビューのポジティブな意見、あるいは、ネガティブな意見でさえも宿泊施設へ注目を集める効果があり、消費者の意思決定を促進するとされている [2]。加えて同じ研究では、ポジティブな意見は消費者の宿泊施設への態度や行動を改善するために役立つとされており、宿泊レビューは消費者にも提供側にも貴重なリソースと言える。

宿泊レビューにおける消費者の意思決定要因を考える際に、本研究では宿泊レビューで時間的に変化する特徴に着目した。季節によって大きく気候が変わる地

域では、各季節で衣食住の様子や人気の旅行先は異なる。また、近隣で開催される季節ごとのイベントなどによっても各宿泊施設の価値は大きく変化する。宿泊レビューに含まれる季節的な特徴を抽出して提示することで、宿泊検討者は宿泊する予定の季節の情報を元に宿泊先を決定することができる。

本稿では、宿泊施設の各月のレビューから重要度の高い単語を抽出するとともに、その重要単語を含む階層型のトピックを、月間で比較することによって、宿泊施設の季節的な特徴を分析する手法を提案する。テキストマイニングにおいて文書内の名詞に与える代表的な特徴量として $tf-idf$ 法による値があり、文書分類や機械学習の素性として用いられる [3]。本研究では名詞に月別の $tf-idf$ 値を付与することで、季節の特徴語の抽出を行う。宿泊施設の特徴抽出では、消費者の欲しい情報を的確に明示するためには“食事”や“風呂”など、カテゴリーの存在は考慮しなくてはならない [4]。カテゴリーによっては季節に応じて特徴が変化し、消費者の決定要因となるかどうかも変化する。提案手法では、カテゴリーの取得にトピックモデル（潜在的意味解析）を用いる。トピックモデルの一つに LDA (Latent Dirichlet Allocation) [5] という教師なし学習手法がある。この手法では、ある文章は複数のトピックを持つと仮定し、そのトピックを抽出することを目

*連絡先: 〒 525-8577 滋賀県草津市野路東 1-1-1
E-mail: {is0309he@ed, ryama@media, nisi-hara@fc}.ritsume.ac.jp

的としている。宿泊レビューに対し LDA を使った分析に、Han らの先行研究がある [6]。この研究では、宿泊施設レビューのトピック分析により感情と評価の関係を抽出することに成功している。また、LDA の拡張手法に hLDA(hierarchical Latent Dirichlet Allocation) がある [7, 8]。hLDA では文書に含まれるトピックに階層構造が存在すると仮定し、トピックを確率的に推定する技術である。宿泊レビューの特徴について、Wang らの先行研究 [9] では、レビューのトピック分析から潜在的な意見・評価を得る Latent Aspect Rating Analysis (LARA) という問題が定義されている。LARA から、観点の多さは宿泊レビュー特有の特徴とも言える。観点や評価属性を、カテゴリーとして考えると、宿泊レビューのカテゴリーの中には包含関係が存在する。例えば“食事”に関する評価の中には、“食事の値段”や“食事の質”などより詳細な情報が含まれる。本稿では、hLDA を用いた分析によって宿泊レビューのトピックの包含関係が抽出できると仮説を立て、その結果と $tf-idf$ 法により抽出した重要単語を用いた宿泊レビューの時間変化分析を行う。

2 分析データ

本稿では、楽天トラベルのレビュー 5,082,427 件のユーザの声をを用いる¹。使用するレビューデータは 1996 年から 2016 年の間に収録されたものであり、29,400 件の宿泊施設に対してのレビューが存在する。本研究の分析では統計的な見地を導出するため、宿泊施設の中でもレビュー数上位 10% の宿泊施設を分析に使用する。本稿では、このうちレビュー数上位 10% の宿泊施設からランダムに選出した以下の 2 件の宿泊施設を今回の分析対象とした。

1. 粟津温泉 旅亭懐石 のとや
2. シーサイドホテル舞子ピラ神戸

これらの宿泊施設のレビューを月別に分割し、一つの宿泊施設につき 12 個の文書集合を入力データとして扱う。

2.1 分析対象とする単語

自然言語処理の前処理の重要性は広く知られている。ストップワードや分析する内容に応じた品詞の絞り込みは、分析の精度向上につながる。

本稿では、宿泊施設の特徴の抽出をメインとするため、名詞のみを分析対象とする、トピック抽出を目的と

¹楽天株式会社 (2016): 楽天トラベルデータ。国立情報学研究所情報学研究所データリポジトリ。(データセット). <https://doi.org/10.32130/idr.2.2>

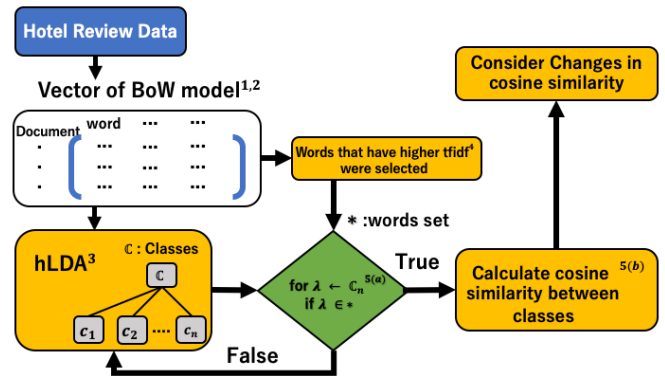


図 1: 分析手順のフレームワーク。図中、3 章に示した分析手順番号を肩文字として示す。

しているため、それ単体で意味のわからない名詞、日にちや記号に関してはストップワードとして扱い、分析から除外する。さらに、hLDA の結果の収束を良くするために、出現回数 3 回以上の名詞のみを使用する。

3 hLDA と $tf-idf$ 法を用いた季節に存在する潜在的なトピック分析

提案手法では、宿泊施設の潜在的なトピック分析として hLDA、季節の特徴抽出に $tf-idf$ 法を用いる。この 2 種類の手法により宿泊レビューに存在する季節の潜在的トピックの階層構造での抽出を行う。時間変化分析として、抽出された季節の特徴が他の月でどのようにクラスタリングされているかを考察し、季節の特徴語の変化、トピックの移り変わりを分析する。抽出と考察は、以下の手順 (図 1 の添字番号を参照) に沿って行う、

1. レビューの分かち書きの際に形態素解析器 MeCab と辞書に NEologd を用いて出現した名詞を取得する。
2. Bag-of-Words モデルに基づき出現した名詞のベクトル化を行う。
3. 作成したベクトルを入力し、hLDA で潜在的なトピックを階層構造でクラスタリングする。
4. 分析する月単位のレビューを一つの文書、他の月のレビューも含め 12 個の文書集合とし、出現した名詞に $tf-idf$ 値を付与する。
5. ストップワードを除く $tf-idf$ 値の上位 10% の名詞について、次の (a) と (b) の処理を行う：
 - (a) 同じクラスに抽出された名詞の傾向を分析する。

- (b) 他の月の（その名詞を含む）クラスタと類似度を計算する。

3.1 階層的トピックモデル

入力データとして用いる宿泊レビューには、大カテゴリと小カテゴリが存在するように、トピックとしても抽出可能であると仮定し、このトピックの階層関係を抽出する。LDAモデルの拡張として、トピックの階層関係を解析できる hierarchical Latent Dirichlet Allocation (hLDA) に着目し、宿泊レビューの季節的なトピックの変化を評価する。

3.1.1 nested Chinese Restaurant Process

nested Chinese Restaurant Process (nCRP) とは木構造上の確率過程である。この確率過程は hLDA のために用いられ、CRP を利用して以下の比喩で表される [10]。

CRP は N 人の客が無限のテーブルを持つ中華レストランに座るプロセスを仮定して得られる分布のことである。レストランに入ってきた順に、 $1, 2, \dots, N$ とラベル付けできる客が、レストランのテーブルを選択して座っていくとする。最初の客は最初のテーブルに座り、 n 人目の客が座るテーブルを c_n とし、 i 番目のテーブルに座る確率は式 (1) によって決定される。

$$p(c_n = i | c_{n-1}) = \begin{cases} \frac{n_i}{\gamma + n - 1} & (\text{occupied table } i) \\ \frac{\gamma}{\gamma + n - 1} & (\text{next unoccupied table}) \end{cases} \quad (1)$$

ここでの n_i は現在テーブル i に座っている顧客の数であり、 γ はレストランの顧客の数と比較して、顧客が新しいテーブルを選択する頻度を制御する実際のパラメータである。

nCRP において、客は CRP に基づき木構造を形成する。そこで nCRP は以下の比喩で表される。街に無限個のレストランが存在し、各レストランは無限個のテーブルを持つと仮定する。また階層の根となるレストランが存在し、各レストランのテーブルは他のレストランを指定している。まず最初の客が根のレストランに入り、CRP に従いテーブルを選択し、次のレストランへのルートを得る。再び CRP に従いテーブルを選択し、これを無限回繰り返す、木構造におけるパスを決定する。これをすべての客について行うと、無限に枝分かれした無限深さの木の部分木を得る。本研究では客が単語、レストランが階層、テーブルの席がトピックを表し、この部分木の決定に用いられる。

3.1.2 hLDA の生成過程

hLDA の生成過程 [10] において、木構造は nCRP によって生成され、nCRP における客が文書、テーブル（またはレストラン）がトピックを表している。ここからは、先行研究 [7, 8] にならい、hLDA の生成過程を以下に記す。

1. 各トピック $k \in T$ に対して
 - (a) 単語上の多項分布パラメータ $\beta_k \sim Dir(\eta)$ を選択
2. 各文書 $d \in \{1, \dots, D\}$ に対して
 - (a) レストランプロセスに基づきノードを選択 $c_d \sim nCRP(\gamma)$
 - (b) レベル上の多項分布パラメータ $\theta_d \sim Dir(\alpha)$ を選択
 - (c) 各単語 $w_{d,n} (n \in \{1, \dots, N_d\})$ に対して
 - i. レベル $z_{d,n} \sim Mult(\theta_d)$ を選択
 - ii. 単語 $w_{d,n} \sim Mult(\beta_{c_d, z_{d,n}})$ を選択

hLDA の実装では、あらかじめパラメータ (α, γ, η , 階層数) を設定する必要がある。このハイパーパラメータ問題に対し、分析対象の宿泊施設によって抽出の「適切さ」が異なる。このことから、同一の条件で抽出と「適切さ」を得るために先行研究 [7] にならい、より多くのトピックを抽出するため、確率分布の中レベルの層が抽出されやすいように、 $\gamma = 1.0, \eta = 1.0$ に定めた。

本稿では、宿泊施設の特徴の抽出を大きな目標としているため、そのトピックが何に基づいて分類されたかがわかる方が望ましい。そこで本研究では宿泊施設の特徴のトピックの階層数を、よりトピックの内容が特徴的である階層数 3 に設定し、収束した名詞のみを分析に用いた。

3.2 $tf - idf$ 法に基づく季節の特徴分析

$tf - idf$ 値は、その文書内で使われている単語の重要度を表す [3]。対象の文書内での出現頻度が高い単語ほど重要であるという考えに基づく tf 値と、いくつもの文書で横断的に出現する単語は重要でないという考えに基づく逆文書頻度、 idf 値の掛け合わせで $tf - idf$ 値を計算する。月毎のレビューを文書集合として入力するため、 $tf - idf$ 値の高い名詞は、他の月に出現していない季節を表す特徴的な名詞と言える。 $tf - idf$ 値の特徴量付与は tf 値を式 (2)、 idf 値を式 (3) で行う。

$$tf(t, d, month) = \frac{n_{t,d,month}}{\sum_{s \in d} n_{s,d,month}}, \quad (2)$$

$$idf(t, month) = \log \frac{N}{df(t, month)}. \quad (3)$$

3.3 他クラスタ間コサイン類似度

一般的に自然言語処理で使われるコサイン類似度は、ベクトル化した文書間の類似性を測ることによく使われる。本稿では、hLDAによって分類された月別のクラスタの周辺単語の違いから、その季節を特徴付ける名詞や特徴の移り変わりを抽出する。そこで、hLDAで獲得された各月のクラスタ間でコサイン類似度を計算する。

クラスタは複数の名詞の集まりで形成され、それぞれがBoWモデルに従いベクトル化した値を計算に用いる。m1月とm2月のクラスタC_{m1}とC_{m2}の集合を、式(4)と式(5)に記す。

$$C_{m1}(1, \dots, n) = \{W_1, W_2, \dots, W_n\}, \quad (4)$$

$$C_{m2}(1, \dots, n) = \{W_1, W_1, \dots, W_n\}. \quad (5)$$

この2つの集合に対し、コサイン類似度の計算は式(6)で行う。

$$\begin{aligned} \cos(C_{m1}, C_{m2}) &= \frac{\sum_{k=1}^n C_{m1}(k) \cdot C_{m2}(k)}{\sqrt{\sum_{k=1}^n (C_{m1}(k))^2} \cdot \sqrt{\sum_{k=1}^n (C_{m2}(k))^2}}. \end{aligned} \quad (6)$$

4 分析結果

ユーザのニーズの多さという観点から分析対象を宿泊施設の繁茂期に絞り、1月、5月、8月、12月の結果を出力した。2件の宿泊レビューの名詞のうち、tf-idf値の高かった上位5件を表1に記す。

図2、図3はhLDAの取束しない名詞を除いた分析結果の一部である。次に表1から、季節そのものや時期を表す名詞以外の“食事”や“イベント”など評価の対象として相応しいものを選択する。本稿では「粟津温泉 旅亭懐石 のとや」の“鯛”, 「シーサイドホテル舞子ピラ神戸」の“プール”に着目する。また分析するクラスタを、以下のように定義する。

小クラスタ 該当する名詞を含む階層2の単一のクラスタ、図2、図3に表す所のC_{a1}やC_{a2}に当たる。

大クラスタ 該当する名詞を含む階層3のクラスタと同じ親を持つ複数のクラスタ、図2、図3に表す所のC_aやC_bと、それぞれの子クラスタ全ての集合。

表 1: 2件の宿泊施設のレビューから抽出したtf-idf値が高かった名詞, 上位5件

	粟津温泉 旅亭懐石 のとや	シーサイドホテル舞子ピラ神戸
1月	お正月, 年越し, 年始年末, 餅搗ぎ, 最上級	お正月, お年玉, お菓子, スペシャル, 周年
5月	ゴールデンウィーク, 鯛, 春の, 愛, 連休	室温, ゴールデンウィーク, 週末, グレード, エステ
8月	お盆, 合掌, 道場六三郎, 鮎, 納涼	プール, 夏休み, 海水浴場, 海水浴, バーベキュー
12月	クリスマス, めし, 突破, 名物, 最上級	ルミナリエ, クリスマス, 冬, おもてなし, スペシャル

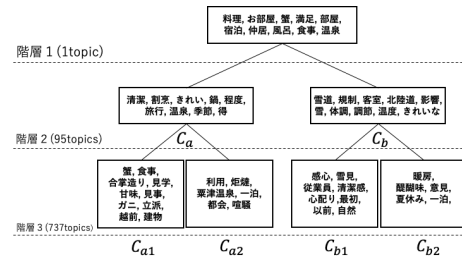


図 2: hLDA, 1月, 粟津温泉 旅亭懐石 のとや

このように大クラスタと小クラスタの類似度の推移を分析することにより、トピックの時間的な大規模な変化と小規模な変化を比較することが可能になる。この結果を表2、表3にそれぞれ記す。また注目する名詞を含む他の月の階層構造を図4、図5に示す。

5 考察

5.1 hLDAの抽出結果

図2のC_aは“清潔”や“温泉”, “鯛”の一般的なサービスや、施設への評価がトピックとして存在することがわかる。対象的にC_bには“雪道”や“雪”, “温度”など天候や環境に関するトピックが抽出された。またC_{a1}には“蟹”に関する名詞や“甘味”といった食事に関する評価が抽出され、C_{a2}~C_{b2}まで比較しても明らかに食事関連のトピックが分類されていることがわかる。またC_{b1}はサービスや環境に対してのトピック、C_{b2}は温度に対してのトピックであり、明確な差が見取れる。ここからC_bとC_{b1}, C_{b2}の階層関係が見取れる。

図3のC_aは価格、立地に関するトピックが存在することが見てわかる。他にC_bには“海水浴場”や“眺め”から外のレジャー、あるいは施設の特徴がトピックとして存在することがわかる。C_{a1}, C_{b1}に、“家族連れ”や“海水浴場”が分類されており、家族旅行が楽しめ、海水浴ができることが想像される。実際にそのようなレ

表 2: 粟津温泉 旅亭懐石 のとや : $tf-idf$ 法で抽出された 5 月の “鯛” を含む他クラスタ間のコサイン類似度. 表中, 同クラスタ間の類似度を下線部に示す.

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
大クラスタ	0.0	0.0	0.0	0.029	<u>1.0</u>	0.046	0.018	0.039	0.021	0.028	0.0	0.0
小クラスタ	0.0	0.0	0.0	0.099	<u>1.0</u>	0.199	0.099	0.099	0.099	0.105	0.0	0.0

表 3: シーサイドホテル舞子ピラ神戸 : $tf-idf$ 法で抽出された 8 月の “プール” を含む他クラスタ間のコサイン類似度. 表中, 同クラスタ間の類似度を下線部に示す.

	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
大クラスタ	0.0	0.0	0.0	0.0	0.0	0.110	0.042	<u>1.0</u>	0.164	0.0	0.0	0.0
小クラスタ	0.0	0.0	0.0	0.0	0.0	0.099	0.099	<u>1.0</u>	0.099	0.0	0.0	0.0

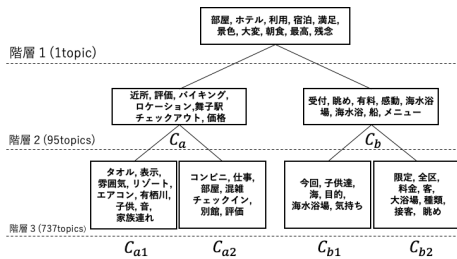


図 3: hLDA, 8 月, シーサイドホテル舞子ピラ神戸

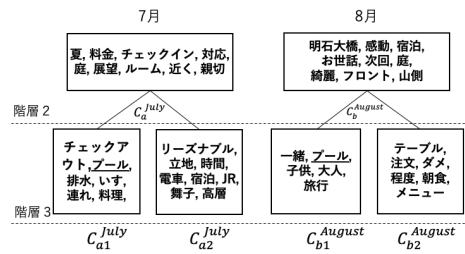


図 5: “プール” を含む 7 月, 8 月の階層構造

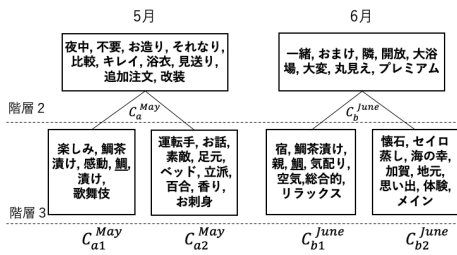


図 4: “鯛” を含む 5 月, 6 月の階層構造

ビューも存在した. これも hLDA の結果の特徴であると考え.

結果として, [粟津温泉 旅亭懐石 のとや] の宿泊レビューから, 1 月に料理, 特に蟹や鍋が楽しめ, また天候では雪が想定されることがわかる. また [シーサイドホテル舞子ピラ神戸] の 8 月の宿泊レビューから, 家族連れが海水浴に訪れやすい季節であることがわかった.

5.2 コサイン類似度による他クラスタ間評価

表 2 の結果から, 大クラスタの推移を見ると, 6 月が一番高く, 次に 8 月, 4 月が高いことがわかった. 次元数が増えることも影響し, 大クラスタでさまざまなト

ピックを含むため, より詳細な変化が見て取れる. また, 5 月に対して, 4 月と 6 月の類似度が高いことから, 近いトピックが存在することがわかる. このことから季節の類似性は宿泊レビューのトピックに現れることがわかる. 図 4 の “鯛” を含む 5 月, 6 月の階層構造には, “鯛茶漬” がどちらの 3 階層にも含まれており, 料理名が共起している. 実際には鯛の旬は 3 月~6 月と 9 月~11 月と言われている, 旬を考慮しなければオールシーズン食べることができる魚である. しかし, この宿泊施設の場合, 5 月 6 月で注目されているトピックであることが明らかになった.

表 3 の結果から, “プール” は 6 月から 9 月まで出現していることがわかる. 大クラスタにおいて, 類似度は 9 月が一番高くなっており, 近い季節において似たトピックが存在することがわかる. しかし, 一般的な “プール” の人気が高まる 7 月と 8 月の類似度が低くなった. 図 5 から, “プール” と同じクラスタに “大人” や “子供” といった家族旅行の様子がわかる. 実際のレビューにも, 「家族連れ」の旅行形式が, 7 月より 8 月に多く出現する. この結果から, 家族旅行に適した季節, またはそうでない季節が類似度の結果から明らかになった.

結果から, 季節で注目される食事内容や, 家族旅行に適した季節, または家族連れで賑わう季節が明らか

になった。

5.3 今後の展望

改善点・展望として、(1)hLDA のハイパーパラメータに起因する精度問題、(2) 意思決定支援に有効な可視化手法、(3) 地域差に現れるトピックの差の抽出、が挙げられる。それぞれ、(1) については分析精度の向上、(2) と (3) については宿泊検討者の意思決定支援に有効なインタラクションモデルに関わる問題と分類される。

hLDA の結果として分類が収束しない名詞も見られ、精度の向上が必要なことは明らかである。解決策としてハイパーパラメータの調節が有効であるが、宿泊施設の各文書によって調整を行なうことは、手動では限界があり、可視化手法の発展が必要である。本稿では、分析と結果の考察までを紹介したが、大きな目標である“消費者の意思決定支援”につなげるこの本質を見抜く必要がある。そこで、決定要因になり得る情報の選別などが発展として期待できる。また季節と同じように地域差でも特徴量の違いを抽出できることが予想される。

6 おわりに

本稿では、宿泊検討者に対して意思決定支援につながる分析を目的として、季節毎に重要な単語の意味的な変化を抽出する手法を提案した。提案手法では、hLDA の分析によって宿泊施設の大カテゴリーとその中にある小カテゴリーを階層構造で抽出することを試みた。このとき、精度の定量的な評価や、注目すべき名詞の選別として、コサイン類似度と $tf-idf$ 法を用いて分析した。その結果、宿泊施設のトピックを階層関係で抽出でき、季節特有のトピックの情報を抽出できた。また、ハイパーパラメータに起因する精度の問題が明らかになった。今後の展望として、季節と同じように地域差でも特徴量の違いを抽出でき、宿泊検討者への可視化を行う予定である。

謝辞

本研究では、国立情報学研究所の IDR データセット提供サービスにより楽天株式会社から提供を受けた「楽天データセット」を利用した。記して謝意を表す。

参考文献

[1] Astrid Dickinger and Josef Mazanec. Consumers' preferred criteria for hotel online booking. *In-*

formation and communication technologies in tourism 2008, pp. 244–254, 2008.

- [2] Ivar E Vermeulen and Daphne Seegers. Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism management*, Vol. 30, No. 1, pp. 123–127, 2009.
- [3] Juan Ramos, et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242, pp. 133–142. Piscataway, NJ, 2003.
- [4] Katerina Berezina, Anil Bilgihan, Cihan Cobanoglu, and Fevzi Okumus. Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, Vol. 25, No. 1, pp. 1–24, 2016.
- [5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.
- [6] Hyun Jeong Han, Shawn Mankad, Nagesh Gavirneni, Rohit Verma, et al. What guests really think of your hotel: Text analytics of online customer reviews. 2016.
- [7] David M Blei, Thomas L Griffiths, and Michael I Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, Vol. 57, No. 2, p. 7, 2010.
- [8] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*, pp. 17–24, 2004.
- [9] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 783–792, 2010.
- [10] 山本浩平, 江口浩二, 高須淳宏. カテゴリ階層の拡張を目的とした階層的トピックモデル. 第 6 回データ工学と情報マネジメントに関するフォーラム, 2014.

分析対象と目的に基づき データ分析手法の選択を支援するインタフェースの検討

Supporting Interface for Choosing Text Mining Methods of TETDM Designed for Analysis Unit and Purpose

西原陽子^{1*} 深井俊樹¹ 山西良典¹
Yoko Nishihara¹ Toshiki Fukai¹ Ryosuke Yamanishi¹

¹ 立命館大学情報理工学部

¹ College of Information Science and Engineering, Ritsumeikan University

Abstract: Text data analysis can be conducted by using text mining software. Instructions and tutorials of the software tell users how to use the software. TETDM, one of the text mining software, also has tutorials for users. Even if the users are the beginners of text mining or data analysis, they would be able to use TETDM after finishing the tutorials. However, the users need much time to finish all of the tutorials because there are many quizzes that can be solved by using tools of TETDM. Moreover, though the users can do text mining by referring the quizzes of tutorial, they also need much time to find quizzes that are related to their own analysis unit and purpose. We believe that it is necessary to support users for finding appropriate tools in shorter period. This paper proposes a supporting interface for choosing text mining methods of TETDM considering analysis unit and purpose. Once a user decides the analysis unit and purpose of text mining, the interface shows appropriate methods of TETDM. We had evaluation experiments with the proposed interface. We asked participants to answer questions of text mining by using the proposed interface. Experimental results showed that the participants answered questions more correctly and in shorter periods.

1 はじめに

多くの分野で収集されたデータを分析し、業務に活用しようとする動きが進められている。データの分析が手作業ではなく、パソコン上のツールを用いて行われることも増えてきた。大学においてもデータサイエンスの講義や演習が開講されるようになり、ツールを用いたデータ分析のスキルを身につけることの需要が高まってきている。

パソコン上のツールを用いてデータ分析を行うと、大量のデータを短時間で処理することができるようになる。しかし、データ分析を行うためには、目的や対象に対して適切なツールを選択し、ツールを使いこなす必要がある。仮に分析する人がデータ分析の初心者であると、自分の分析したい対象や目的にとって最適なツールがどれかが分からず、分析が進められないこともあると考えられる。

パソコン上のツールには説明書やチュートリアルが付

属していることがある。ツールの説明書の多くは、ツールの使い方を説明してくれ、チュートリアルも多くはサンプルデータを用いて、ツールの分析結果の見方などを説明してくれる [3]。

テキストマイニングのソフトウェアである TETDM にもチュートリアルがあり、スーパーライトモード、ライトモード、通常モード、拡張モードと 4 種類のチュートリアルが用意されている。テキストマイニングやデータ分析の初心者であっても、チュートリアルを全てクリアすると、TETDM が使えるようになるとされている。しかし、チュートリアルの問題数は多く、全てをクリアするまでに多くの時間がかかってしまう。また、チュートリアルを参考にして、TETDM のツールを使うおうとすると、ツールの使用が説明されているチュートリアルの箇所を探す必要があり、時間がかかってしまう。ユーザが分析したい対象と目的を持ち、ツールを使いたいと思っても、その説明がある場所を探すことが大変になっている。より短時間でツールを探せる仕組みが必要と考えられる。

そこで本研究では、ユーザの分析したい対象と目的に

*連絡先：立命館大学情報理工学部
滋賀県草津市野路東 1-1-1
E-mail:nishihara@fc.ritsumeikai.ac.jp

応じて、テキストマイニング手法の選択を支援するインタフェースを提案する。本研究ではテキストマイニングのソフトウェアであるTETDMのツールの選択を支援することとする。分析したい対象は、本研究では単語、文、段落、文章の4つのうちのいずれかとする。選択肢として提示するテキストマイニング手法は、TETDMに含まれている手法のいずれかとする。TETDMのチュートリアルを利用する場合と、提案インタフェースを利用する場合を比較し、提案インタフェースにより短時間で分析ツールにたどり着け、分析が行えることを目標とする。

多くの選択を支援するシステム、ホテルの検索システム(例えば、<https://www.booking.com/>)やレストランの検索システム(例えば、<https://www.gnavi.co.jp>)などにおいては、ユーザが必要とする条件を入力させることにより選択肢を絞り込み、絞り込んだ選択肢をユーザに提示している。本研究でも同様に、ユーザが必要とする条件として分析したい対象と目的の2つを入力させて、利用可能なツールを絞り込みユーザに提示することとする。

ユーザの嗜好を捉えた上で選択支援を試みる研究も多数行われている。階層分析法(AHP)を用いた研究[1, 4]や、ユーザの嗜好の点数付けから評価軸を作成する研究[2]などもある。本研究ではデータ分析の初心者のためのツール選択支援システムについて検討をする。初心者はツールを熟知しておらず、熟知していない状態ではツールに対する嗜好は存在しないと考えられる。したがって、嗜好を用いた選択支援については対象とせず、データ分析に必要な条件が入力されたときの選択支援を対象とする。

2 提案インタフェース

提案インタフェースの概要を図1に示す。提案インタフェースは、サイドメニュー、メイン画面の2つから構成される。サイドメニューの一番上には分析対象である「単語」「文」「段落」「文章」の4つが示されている。ユーザが分析したい対象を選択すると、図2に示すように、分析目的が示される。分析目的を選択すると、サイドバーに利用可能な手法が表示される。ユーザが手法を選択すると、手法に必要なTETDMのツールにポップアップがつけられる。ポップアップをクリックすると、図3に示すように、手法の実行例が表示される。提案インタフェースを用いることにより、ユーザは分析対象と目的に合致する手法を選択していくことが可能になる。



図1: 提案インタフェース概要。分析対象や目的を表示するサイドメニュー(左)と、対象や目的に利用可能なツールを表示するメイン画面(右)の2つにより構成される。



図2: 分析目的の表示例。図では対象として「文」が選択された時に、目的として「特定の単語の抽出」「単語間の関係の抽出」「単語の情報の抽出」「その他」が表示されている。



図 3: 手法の実行例の表示例. 図 1 のメイン画面に表示されているポップアップをクリックすると、手法の実行例が表示される。

2.1 分析手法の分類方法

ユーザがテキストマイニングの対象と目的を入力した時に、利用可能な手法を表示するため、TETDM内の手法を分類する。著者は最初に、TETDMに含まれる手法を分析対象により分類し、その後分析目的により分類を行った。本研究における分析対象は単語、文、段落、文章の4つになり、TETDMに含まれる手法をまず4通りに分類した。続いて、分析目的により詳細な分類をした。本研究における分析目的は対象の属性、対象間の関係、対象が持つ情報の3つになる。分析対象と分析目的による分類クラスを表1の左2列に示す。

TETDMではデータを分析する際に、分析手法と分析結果の可視化手法と2つの手法を設定する必要がある。分析目的ごとに手法のペアが設定されており、分析手法と可視化手法を分けて分類することが困難であった。そこで、本研究では分析手法にだけ注目し、分析対象と目的を考慮して分析手法のみを分類した。分析手法のペアとなる可視化手法は、分析手法が分類された先に自動的に振り分けた。

手法の分類結果を表1に示す。分析対象ごとに分析手法を分類し、その後、分析目的ごとに分析手法を分類した。共通する分析目的が存在しなかった分析手法については、その他というクラスを設け、そこに分類した。

3 評価実験

提案インタフェースによる選択支援について評価する実験を行なった。実験は以下の手順で行なった。

表 1: 手法の分類結果

対象	目的 (抽出物)	手法の種類数 (例)
単語	特定の単語	9 (専門用語抽出など)
	単語間の関係	2 (単語間関連度など)
	単語の情報	7 (単語情報まとめなど)
	その他	3 (辞書オンラインなど)
文	特定の文	6 (意見文抽出など)
	文間の関係	0
	文の情報	1 (文情報まとめ)
	その他	1 (テキスト集合評価アプリ)
段落	特定の段落	1 (テキスト分類)
	段落間の関係	6 (段落順序評価など)
	段落の情報	2 (セグメント情報まとめ)
	その他	1 (テキスト集合評価アプリ)
文章	特定の文章	0
	文章間の関係	0
	文章の情報	5 (テキスト評価など)
	その他	3 (テキスト評価アプリなど)

1. 実験者は被験者を実験群と統制群に分ける
2. 各群の被験者は、TETDMのチュートリアルを2日間に渡って行う
3. 各群の被験者は、実験者が出題するテキストマイニングの問題に回答する

実験者は第2著者であった。被験者は情報理工学部に所属する大学生14名であった。実験群7名、統制群7名であった。TETDMを利用した経験はなく、データ分析の経験も少ない被験者であった。

TETDMのチュートリアルは4種類あり、それぞれスーパーライト、ライト、通常、拡張モードとなっている。1日目にスーパーライトとライトモードを行い、2日目に通常と拡張モードを行なった。チュートリアルにはツールの説明以外にも、データの入力方法やTETDMの機能についても説明があり、TETDMを用いたデータ分析を行うためにはチュートリアルを行うことが必要であった。そのため、実験群と統制群の両群がチュートリアルをこなした。

手順の3.で問題に解答する際、実験群は提案インタフェースを用いた。統制群はTETDMのみを用いた。

テキストマイニングに用いた文章を表2に示す。学生が興味を持ちやすいタイトルで、文字数が約1000ずつ異なる文章を7本用意し、実験を行なった。

出題した問題を表3に示す。これらの問題はテキストマイニングの一般的な問題として用意した。

表 2: 実験で用いた文章のタイトル

番号	タイトル	文字数
1	あおり運転の心理	2118
2	ガンにならないのはどんな人か?	3177
3	「すぐに返信しない男」と「既読スルーを我慢できない女」の脳の違い	4323
4	ゲームのやりすぎは本当に「精神障害」なのだろうか?	4470
5	いつまでも消えないタバコ「7つの神話」の真実を明かそう	7626
6	ウォーキング・デッドのストーリーまとめ	9503
7	ハリーポッターシリーズのストーリーまとめ	21068

表 3: 実験で出題した問題

問題番号	問題文
1	最も出現頻度が高い単語は何か? 単語名をこたえよ.
2	第3段落で tfidf 値が1番高い単語はどれか? 単語名をこたえよ.
3	関連度がT以上のとき, 単語Xと関連のある単語はいくつあるか? 単語の個数を答えよ.
4	最も独自性の高い段落は何段落目か? 段落数をこたえよ.
5	単語Yに関わっている段落はいくつあるか? 段落の個数を答えよ.
6	最も複数の段落と関連(類似)しているのは何段落目か? 段落数を答えよ.
7	主題に関連した文は文章中(前半, 後半)どちらに多く登場するか?

表 4: 問題の正解率

問題番号	実験群	統制群
1	100%	100%
2	71%	85%
3	57%	0%
4	86%	57%
5	71%	42%
6	28%	0%
7	100%	85%
平均	73%	53%

表 5: 問題の解答時間

問題番号	実験群	統制群
1	58 秒	68 秒
2	98 秒	78 秒
3	515 秒	711 秒
4	201 秒	265 秒
5	252 秒	642 秒
6	361 秒	711 秒
7	164 秒	231 秒
平均	236 秒	387 秒

3.1 実験結果

問題の正答率を実験群と統制群に分けて表4に示す. 平均の正答率は, 実験群が73%, 統制群が53%であり, 実験群のほうが正答率が高かった.

続いて, 問題の平均解答時間を実験群と統制群に分けて表5に示す. 平均の解答時間は, 実験群が236秒, 統制群が387秒で, 実験群のほうが解答時間が短かった ($t=-4.84, p=0.0008 < 0.01$).

3.2 考察

問題の正答率は実験群の方が統制群よりも高かった. 実験群は分析する単位と目的を問題文から読み取り, 適切な手法を選択できたため, 正答率が高くなったと考えられる.

回答時間は実験群の方が統制群よりも短かった. 実験群は適切な手法を選択でき, 解答が得やすかったため, 回答時間が短くなったと考えられる.

以上の結果から, TETDMでのツール選択に提案インタフェースを利用することで, テキストマイニングの問題が正しく解け, さらに短時間で解けることが示された.

4 おわりに

本研究では、テキストマイニングのソフトウェアである TETDM を用いてデータ分析をするときに、手法の選択を支援するためのインタフェースを提案した。提案インタフェースでは、ユーザが分析したい対象と目的を有した時に、それに合致する手法を表示し、選択の支援を行うものである。評価実験を行ったところ、TETDM を使用するとき提案インタフェースを用いることにより、テキストマイニングの問題の正答率が高くなり、また回答時間も短くなることが確認された。本研究では TETDM を対象として研究を進めたが、高度なソフトウェア、システムのユーザを支援する方法についての研究であり、類似する他のソフトウェアにも応用できる知見が得られたと考えている。チュートリアルを解くことにより、TETDM の使い方を一通り学ぶことができるというのは変わらず、今後は提案インタフェースと TETDM のチュートリアルを併用し、データ分析ができる環境づくりを進めていきたい。

謝辞

本研究の一部は科研費（16K00307）の補助を受けて行われました。記して謝意を申し上げます。

参考文献

- [1] 井上拓子, 原田利宣, ファジィAHP を用いた製品レコメンドシステムの開発, 日本感性工学会論文誌, Vol.11, No.2, pp.255-263, (2012).
- [2] 西原陽子, 赤井れい子, 砂山渡, 橘啓八郎, 積極的思考支援のためのキーワード選好インタフェース, Vol.18, No.5, pp.766-776, (2007).
- [3] 西原陽子, 中垣内李菜, 川本佳代, 砂山渡, TETDM を用いたテキストマイニングのスキル獲得を支援するためのチュートリアルシステムの開発, 知能と情報, Vol.27, No.5, pp.771-783, (2015).
- [4] 湯本真樹, 定性的評価にラフ集合の決定ルールを用いた AHP による商品選択支援システムの開発, 電気学会論文誌C, Vol.139, No.9, pp.1080-1091, (2019).

Creating Reverse Dictionary of English Idiomatic Expressions by Mapping Word Embeddings to Singular Vectors

Xiaodong Liu, Rafal Rzepka, Kenji Araki

Faculty of Information Science and Technology

Hokkaido University

Sapporo, Japan

{xiaodongliu, rzepka, araki}@ist.hokudai.ac.jp

Abstract

This paper demonstrates a reverse dictionary that can return relevant idiomatic expressions based on queries (input descriptions) given by users. The implementation of the system can be achieved with a Vector Space Model (VSM). However, when it comes to VSM, although its performance on queries is high under the condition that pairs of common keywords are shared between a query and documents, the limitations such as lack of common words, and information loss after matrix factorization inevitably worsen the performance. To address these limitations, we employed a feed-forward neural network to map between rich literal semantics (word embeddings) and latent semantics (singular vectors) via general literal semantics (Bag-of-Words). A comparatively high performance on queries is achieved by this approach, which can address the limitations to some extent. Therefore, to take the advantages of both the VSM and the neural network, we sorted their outputs (cosine similarity between a query and vector representations of idiomatic expressions) in decreasing order, and significantly improved the performance on queries.

1 Introduction

According to the survey written by (Turney and Pantel, 2010), the VSM was developed for the SMART information retrieval system (Salton and others, 1971), which pioneered many of the concepts that are used in modern search engines (Schütze et al., 2008). One form of VSM is term-document matrix, which contains the documents representing phrases, sentences, or texts, and terms representing all the single words of

those documents. The key idea of VSM is to represent all the documents in a term-document matrix as points in a space (distributional representations). The points that are close in the space can be considered as semantically similar, whereas the points that are far apart from each other are semantically distant. A user's query (pseudo-document) is also represented as a point in the same space as the documents. The documents are sorted in order of decreasing semantic similarity from the query, and then presented to the user.

The original elements in the vector representation of each document are Bag-of-Words. The frequencies of words in a document, to some extent, can determine how relevant the document is to a query. The Bag-of-Words hypothesis is the basis for applying the VSM to information retrieval (Salton et al., 1975). However, the original term-document matrix has some limitations when applied to information retrieval tasks. When the size of the matrix ($M \times N$) is enormous, this will consume a considerable memory footprint. Moreover, the documents used in our experiment are short texts such as phrases and sentences, which makes the matrix sparse. To address this limitation, matrix factorization was introduced, and one standard way to perform this mathematical operation is Singular Vector Decomposition (SVD) (Deerwester et al., 1990), which is engaged in the process of LSA (Latent Semantic Analysis) (Landauer and Dumais, 1997). SVD decomposes the matrix X ($M \times N$) into the product of three matrices U ($M \times K$) $\cdot \Sigma$ ($K \times K$) $\cdot V^T$ ($K \times N$), in which U is a left singular matrix, Σ is a diagonal matrix of singular values, V^T is a right singular matrix, where $K \ll M$ or N . From the perspective of mathematics, this operation is a kind of dimension reduction. On the other hand, from the perspective of computational linguistics, we consider it as a latent semantics detection: U is composed of latent semantics of all words, and V^T is composed of la-

tent semantics of all documents. Proximity in the induced latent space has been shown to correlate with semantic similarity (Mihalcea et al., 2006). Within the right singular matrix, each document has k latent semantics, and then we just need to use some similarity mechanism like cosine similarity to judge which documents are semantically similar to a given query, and then display them to users. Some researchers argue that important information can be lost after matrix factorization (Ji and Eisenstein, 2013). This is one of the limitations that we want to address. We do not consider the latent semantics only, instead we bind it to literal semantics in order to realize a conceptual mapping mechanism. The details are presented in section 5.

Below is an example that shows another limitation of the VSM. Both descriptions are excerpted from two commercial dictionaries, which describe the meaning of the idiomatic expression: “have your cake and eat it too”.

D1 “to get the benefits of two different situations or things when you should only get the benefit of one of them” (Collins Online Dictionary¹)

D2 “to have the advantages of something without its disadvantages” (Oxford Learner’s Dictionary²)

Since they describe the same idiomatic expression, they can be identified as paraphrase. However, common words are barely shared between them, and ‘advantage’ to ‘benefit’ and ‘get’ to ‘have’ do not share the same lemma. Therefore, it would be problematic for VSM to retrieve D2 when D1 is a query. We call this limitation lack of common words.

2 Reverse Dictionary

A reverse dictionary (Sierra, 2000), also known as an inverse dictionary, or search-by-concept dictionary (Calvo et al., 2016) is a system that returns words based on user descriptions or definitions (Zock and Bilac, 2004). For example, given input query “a large aggressive animal with wings and a long tail, that can breathe out fire”, a reverse dictionary would return ‘dragon’ as primary candidate to users. A successful implementation of a

¹<https://www.collinsdictionary.com/>

²<https://www.oxfordlearnersdictionaries.com/definition/english/>

reverse dictionary was proposed in 2016 (Hill et al., 2016) who used neural language embedding model to map dictionary definitions (phrases) to (lexical) representations of the words defined by those definitions, the performance of which could almost rival a commercial online reverse dictionary³.

As defined above, a reverse dictionary of English idiomatic expressions is a system that returns idiomatic expressions to users, given a query of input description. As of writing, to the author’s best knowledge, there is only one online commercial system⁴ for looking up relevant idiomatic expressions given descriptions as queries. We also compared its performance on our test dataset. The implementation of the reverse dictionary of idiomatic expressions can be completed by VSM. However, to address the aforementioned limitations of VSM, our proposed method is presented in this paper.

3 Dataset

We have collected 1,404 common English idiomatic expressions and their corresponding descriptions (definitions) from three online resources⁵. In addition to these, we also added the definitions from Oxford Learner’s Dictionary to the idiomatic expressions in our dataset. As a result, 2,330 descriptions have been collected. While some idiomatic expressions have only one description, others have two or more descriptions. For example, if the idiomatic expression w has descriptions $\{d_1 \dots d_n\}$, then we include all pairs $(w, d_1) \dots (w, d_n)$ as training examples. For the test dataset, we collected 70 input descriptions from Collins Online Dictionary as unseen data describing 70 idiomatic expressions randomly-chosen from our training dataset.

4 Baseline Method

LDA (Latent Dirichlet Allocation) (Blei et al., 2003) is a classic probabilistic model used to build up document-topics distribution and topic-terms distribution. In our experiment, only document-topics distribution was utilized. Since in our task of providing idiomatic expressions for input descriptions, texts are short phrases and sentences,

³<https://www.onelook.com>

⁴<http://www.idioms4you.com>

⁵<https://www.ef.com/wwen/english-resources/english-idioms/>,
<https://7esl.com/english-idioms/>,
<http://idiomsite.com/>

measuring semantic similarity via topics distribution (100 topics in our baseline because it is same as the number of components of LSA set in the proposed method) would not be accurate. The vector representation (see Equation 1) of each idiomatic expression \vec{I} is the averaged document-topics-distribution vectors \vec{s} of all descriptions belonging to the idiomatic expression.

$$\vec{I} = \frac{\sum_{i=1}^n \vec{s}_i}{n} \quad (1)$$

5 Proposed Method

Our inspiration stems from connotations for a single word; for example, the word ‘professional’ has connotations of skill and excellence. Hence, if another single word like ‘expert’ also has similar connotations, we can consider them as semantically similar. Likewise, if two short texts like phrases or sentences have similar latent semantics, they can also be considered as semantically similar, and then they can be matched.

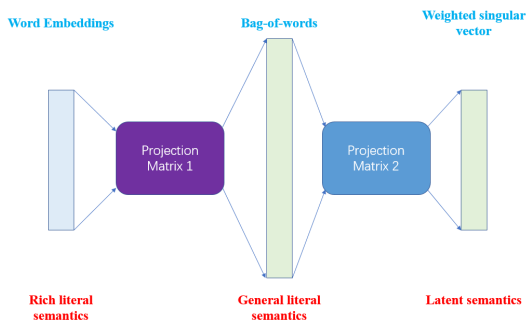


Figure 1: Our feedforward neural network

Figure 1 illustrates the mapping flow of our proposed, the original method described in Section 6. In this section, we focus on the introducing data structure, which is the prerequisite for the training process of the neural network.

The structure of the neural language model can address the limitation of lack of common words. Here is an example of our hypothesis.

- Rich literal semantics: *parents are euphoric about the news of his engagement.*
- General literal semantics: *his parents are excited about his engagement.*
- Latent Semantics: *positive sentiment; marriage*

Although WordNet (Miller, 1998) has well organized taxonomy of words (the word relationship like hyponym to hypernym can project the semantic specificity of words from ‘rich’ to ‘general’), we decided to use GloVe (Pennington et al., 2014), which is considered as an improvement to word2vec model (Mikolov et al., 2013), and trained on billions of words of raw text. The reason that GloVe can outperform WordNet in our case is vector representation of words can be used as the layer in the flowchart of neural network, and elementwise addition of word embeddings; one paradigm is “king – man + woman = queen”. Thus, the elementwise addition of the word embeddings of a short text is considered as rich literal semantics.

As for general literal semantics, in the raw term-document matrix, the 2,330 descriptions have 2,553 single words (with stop-words) and 2,436 single words (without stop-words), and each document (description in this work) is composed of Bag-of-Words. Therefore, the length of each document is 2,553 with stop-words, and 2,436 without stop-words. We use stop-word list provided by the Natural Language Toolkit (NLTK) (Loper and Bird, 2002).

After obtaining general literal semantics, we used “tf-idf reweighting with SVD” (100 components are set in SVD, which is equal to topics number of the baseline) to obtain the singular vectors of the decomposed singular matrix ($2,330 \times 100$). With singular vectors, we can generate the vector representation of all idiomatic expressions. We call the vector representation “averaged singular vector” (namely averaged singular vectors generated by the equation of topics-distribution vectors in the baseline); from the perspective of computational linguistics, we consider all the elements in a vector representation as the latent semantics of the idiomatic expression.

As stated in many papers regarding classification of idioms, e.g. (Peng et al., 2018), idiomatic expressions exhibit the property of non-compositionality. Therefore, we did not use word embeddings or Bag-of-Words to represent idiomatic expressions. In addition, those two cannot represent latent semantics.

6 Implementation Details

Firstly, we trained the Projection Matrix 2 independently from the Projection Matrix 1, in or-

der to map between general literal semantics and latent semantics.

As we mentioned in Section 3, if an idiomatic expression w has n descriptions $(d_1 \dots d_n)$, then we include all pairs $(w, d_1) \dots (w, d_n)$ as training examples. For example, the input of Bag-of-Words d_1 is mapped to averaged-singular-vector(w); the same mapping is performed from d_2 to d_n . After all parameters of Projection Matrix 2 are settled after training, we train Projection Matrix 1.

While Projection Matrix 1 is being trained, the Bag-of-Words Layer is hidden layer with tanh, and the most important thing here is that all parameters in Projection Matrix 2 are not adjusted any more – it is fixed there just like pretrained GloVe. Identically as in the example given above, the input of elementwise-addition(d_1) is mapped to the averaged-singular-vector(w); the same mapping is performed from d_2 to d_n .

We trained two models: with and without stop-words to see if lack of them influence the results. In the case without stop-words, none of the three layers contains any stop-words.

During the test, given a description as an input query, a list of 80 idiomatic expressions is returned in decreasing order of cosine similarity between the output to the query and the averaged singular vectors of all idiomatic expressions.

To compare our model with a conventional VSM, we employed a raw term-document matrix for retrieving idiomatic expressions with and without stop-words; this approach was created to observe the performance change when only literal semantics is engaged. We also implemented LSA with tf-idf (reweighting the term-document matrix using tf-idf and decomposing the matrix using SVD afterwards) to investigate the performance when only latent semantics is utilized.

To combine the advantages of VSM and of our model, we integrated them and evaluated using 70 testing queries. The flowchart of testing a query is illustrated in Figure 2.

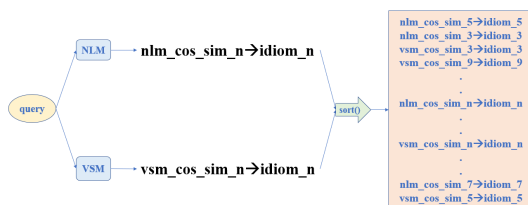


Figure 2: Integration of cosine similarity

The cosine similarities between a model output to a query and the vector representations of all idiomatic expressions are calculated at first, and then sorted into a list according to the decreasing order of cosine similarity.

Items	Settings
Activation Function	tanh
Input Length	200
Output Length	100
Hidden layer w stop-words	2,553
Hidden layer w/o stop-words	2,436
Loss function	squared error
Epochs (Projection Matrix 2)	200
Epochs (Projection Matrix 1)	2,000

Table 1: Neural network hyperparameters

7 Evaluation

Below we describe one observation on the output of our model with stop-words eliminated.

- Input Query: *“a situation of comfort or ease”*
- Idiomatic expression that ranks 4th in the output: *“a bed of roses”*
- The description in the training dataset: *“an easy or a pleasant situation”*

Above is only one observation of the output that ranks at 4th based on the query, which is the desired idiomatic expression that we want to look up. In addition to it, there are the observations that rank at 1st, 2nd or even 50th. Table 2 is not to evaluate but to show the overall performance of 70 queries of test dataset that each model returns.

In the table 2, “@1/10/30/50/80” means out of 70 outputs, what is the percentage that each desired idiomatic expression ranks over top 80, 50, 30, 10, or even exactly at 1.

8 Observation Analysis

Extrapolating from the result of our experiment, we can conclude that VSM performs well only when a query shares many common keywords with target documents; whereas our model can address the limitation of lack of common words. The integration of VSM and our proposed method has outperformed other methods as it takes the advantages of both VSM and our proposed method. Stop-words influence the result of VSM model to

Model	Model Description	@1	@10	@30	@50	@80
Baseline	LDA Topic Modeling	0%	1.49%	2.86%	4.29%	7.14%
LSA + tf-idf (w/o stop-words)	Latent Semantics	4.29%	25.71%	31.43%	38.57%	51.43%
LSA + tf-idf (with stop-words)		10%	35.71%	44.29%	52.86%	57.14%
VSM (w/o stop-words)	General Literal Semantics	37.14%	52.86%	58.57%	58.57%	58.57%
VSM (with stop-words)		32.85%	45.71%	50%	57.14%	58.57%
Proposed method (w/o stop-words)	Proposed Method	28.57%	51.42%	67.14%	72.86%	74.29%
Proposed method (with stop-words)		31.43%	51.42%	65.71%	70%	75.71%
Proposed method + VSM	Integration	38.57%	62.86%	74.29%	80%	82.86%
Idioms4u	Online Commercial	1.43%	20%	27.14%	31.43%	32.86%

Table 2: Comparison results

some extent. On the other hand, it holds almost no influence to our proposed method.

Below we describe three main observations. First, let us show an example that describes the advantage of VSM.

- Input query: “*a speech which is intended to encourage someone to make more effort or feel more confident*”
- The desired idiomatic expression: “*Pep talk*”
- Its descriptions in the training dataset: “*a short speech intended to encourage somebody to work harder*”, and “*an encouraging speech given to a person or group*”
- Baseline: not in the top 80 candidates
- VSM with stop-words involved: 1st
- VSM with stop-words eliminated: 1st
- LSA + tf-idf with stop-words involved: not in the top 80 candidates
- LSA + tf-idf with stop-words eliminated: not in the top 80 candidates
- Proposed method with stop-words involved: 37th
- Proposed method with stop-words eliminated: not in the top 80 candidates

- Proposed method + VSM: 5th
- Idioms4you: not in the top 80 candidates

“Speech”, “intend”, and “encourage” are the common keywords that can contribute to a high cosine similarity between the query and the descriptions. The mapping mechanism worked for one of our models (with stop-words) to some degree, but it did not outperform VSM.

The next example shows how our proposed method is able to address the lack of common words limitation.

- Input query: “*someone who does something or goes somewhere very early, especially very early in the morning.*”
- The desired idiomatic expression: “*an early bird*”
- Its descriptions in the training dataset: “*A person who gets up early in the morning*”, “*A person who starts work earlier than others*”, and “*somebody who does something prior to the usual time*”
- Baseline: not in the top 80 candidates
- VSM with stop-words involved: not in the top 80 candidates
- VSM with stop-words eliminated: not in the top 80 candidates

- LSA + tf-idf with stop-words involved: 44th
- LSA + tf-idf with stop-words eliminated: not in the top 80 candidates
- Proposed method with stop-words involved: 23rd
- Proposed method with stop-words eliminated: 1st
- Proposed method + VSM: 23rd
- Idioms4you: not in the top 80 candidates

Our proposed method outperformed the VSM, even though one of them did not reach top 10. The performance is comparatively good, because the rich literal semantics was to some extent mapped to the general literal semantics. As for VSM, the only common word shared between the long query and the descriptions is ‘early’, so the performance was unsatisfactory.

Finally, the last example below demonstrates poor performance of all models.

- Input query: “*distressed or exasperated to the limit of one’s endurance*”
- The desired idiomatic expression: “*at the end of one’s tether*”
- Its descriptions in the training dataset: “*running out of endurance or patience*”
- Baseline: not in the top 80 candidates
- VSM with stop-words involved: not in the top 80 candidates
- VSM with stop-words eliminated: not in the top 80 candidates
- LSA + tf-idf with stop-words involved: 64th
- LSA + tf-idf with stop-words eliminated: not in the top 80 candidates
- Proposed method with stop-words involved: not in the top 80 candidates
- Proposed method with stop-words eliminated: not in the top 80 candidates
- Proposed method + VSM: not in the top 80 candidates
- Idioms4you: not in the top 80 candidates

In fact, none of the methods worked for this query. Our reasoning is that ‘distress’ and ‘exasperate’ are within the same cluster in the space of word embeddings, but in the description of the training dataset, there exists no general words like ‘sad’ that can be mapped from them. As for VSM, only one common word ‘endurance’ shared between the query and the description in a long phrase, so the cosine similarity between them was very low.

9 Conclusion and Future Work

In this paper, we presented a method for creating reverse dictionary of idiomatic expressions that can return relevant idiomatic expressions given input descriptions as queries. We used feedforward neural network to map between word2vec and singular vectors via Bag-of-Words, and extrapolating from the experimental results, this approach, to some extent, addresses the VSM’s limitation of nonexistent common words. However, not all results of our proposed method achieved superior performance. Besides that, the size of dataset is small, which is not conducive to machine learning. Therefore, in future work, we are going to improve our method by adding idiomatic expressions from Wiktionary and enriching neural network architecture in order to further decreasing information loss after matrix factorization.

References

- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Hiram Calvo, Oscar Méndez, and Marco A Moreno-Armendáriz. 2016. Integrated concept blending with vector space models. *Computer Speech & Language*, 40:79–96.
- Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.
- Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.
- Yangfeng Ji and Jacob Eisenstein. 2013. Discriminative improvements to distributional sentence similarity. In *Proceedings of the 2013 Conference on*

- Empirical Methods in Natural Language Processing*, pages 891–896.
- Thomas K Landauer and Susan T Dumais. 1997. A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.
- Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.
- Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2018. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *arXiv preprint arXiv:1802.09961*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Gerard Salton et al. 1971. The smart system—experiments in automatic document processing.
- Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, page 260.
- Gerardo Sierra. 2000. The onomasiological dictionary: a gap in lexicography. In *Proceedings of the ninth Euralex international congress*, pages 223–235.
- Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.
- Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations: from an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, pages 29–35. Association for Computational Linguistics.

HMMを利用した深層学習ネットワークからの分類パターンの抽出と可視化

Extraction and Visualization of Classification Patterns from Deep Learning Networks using HMM

安藤 雅行^{1,2*} 河原 吉伸^{2,3} 砂山 渡⁴ 畑中 裕司⁴
Masayuki ANDO^{1,2} Yoshinobu KAWAHARA^{2,3} Wataru SUNAYAMA⁴ Yuji HATANAKA⁴

¹ 滋賀県立大学大学院工学研究科

¹ Graduate School of Engineering, The University of Shiga Prefecture

² 理化学研究所革新知能統合研究センター

² RIKEN Center for Advanced Intelligence Project

³ 九州大学 マス・フォア・インダストリ研究所

³ Institute of Mathematics for Industry, Kyushu University

⁴ 滋賀県立大学工学部

⁴ School of Engineering, The University of Shiga Prefecture

Abstract: In deep learning, there is a problem that concrete classification patterns for deriving reasons for classification are often incomprehensible. In this paper, we propose a classification patterns extraction system from deep learning networks and verified the effectiveness of the system. The proposed system extracts classification patterns from the trained learning networks of LSTM using HMM. Then the system displays the extracted classification patterns so that users of the system can interpret the learning networks. In verification experiments, the significance of the extracted classification patterns was estimated by the weights of the classification patterns extracted from data given a unique pattern. The results showed that the proposed system can extract classification patterns effective for interpretations of the learning networks.

1 はじめに

インターネットの普及に伴い、また、SNS (Social Networking Service) の出現によって、画像、テキスト、数値データが大規模になり、その処理や情報の抽出に機械学習が使用されるようになってきた。しかし、従来の機械学習は大量のデータから規則などを学習し、分類・予測を行う際、データのどの特徴（画像なら色や形など）に注目するかは人間が指定する必要があった。そこで注目されるようになってきた技術が、深層学習である。深層学習は近年流行りだした機械学習であり、学習を行う層（入力データの規則などを学習する部分）を多層化している。これにより、より人間の脳の学習に近い段階的な学習ができ、従来の機械学習と比べて学習の精度が高いという利点がある。

一方で、その深層学習による予測・分類基準が人間に

は不明な点が問題になってきている。特に、医療分野や自動運転では、その分類基準の理解は安全性において重要視されている。仮にテキスト分野においても深層学習の判断基準をより深く理解できれば、医療分野において新人とベテランの書いた電子カルテの違いから、良い電子カルテを書く方法を容易に理解でき、企業においても良い報告書や企画書を書く方法を短時間で習得できるなど、深層学習の新しい活用が期待される。

本研究では、構造が複雑になる代わりに、単語の出現の時系列や順序も考慮した学習が可能な、再帰的深層学習（主に LSTM(Long short-term memory)）を使用し、テキスト集合の学習によって構築されたネットワークを HMM(Hidden Markov Model) に当てはめ、ネットワークの層に付けられた重みの値から、入力層に時系列順に入力される特徴量（本研究ではテキストを構成する単語）の尤度を算出する。そして、その単語の順序を考慮した組み合わせを尤度順に取り出すことで、再帰的深層学習の学習済みネットワークに蓄積さ

*連絡先：滋賀県立大学大学院工学研究科 先端工学専攻 安藤雅行
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: oh23mandou@ec.usp.ac.jp

れた情報を、分類パターン（単純な単語の順序列）として抽出することができるシステムを提案する。

以下本論文では、2章で関連研究について述べる。3章でHMMを利用した深層学習による分類パターンの抽出・可視化システムの構成と詳細について述べる。4章で提案システムの評価実験について述べ、5章で本論文を締めくくる。

2 関連研究

インターネットの普及などにより、急速に大規模化しつつあるテキストへの対策として活用され始めているのが、深層学習を用いたテキストマイニングシステムである [1, 2]。深層学習とは、一般に多層から構成されるニューラルネットワークを用いた学習を指し、例えば、深層学習の応用モデルである畳み込みニューラルネットワーク [3] の出現により、画像を用いた場合に限らず多くの場面で高い分類性能を実現できることが報告されている。

その一方で、深層学習は、その出力を導いた根拠についての解釈が困難であることも知られている。画像認識においては、この問題に対する研究も最近進められており、例えば、入力画像に対応する畳み込みニューラルネットワークにおける層間のスコアの勾配を計算することでネットワークの可視化を行う方法 [4] や、学習済みのネットワーク中間層のノード情報を用いて、対応する画像中の画素への寄与度を計算することにより画像の分類に重要な部位を表示する方法などが提案されている [5]。

しかし自然言語への深層学習の適用においては、上記のような画像認識における方法を直接適用できない。そこで、アテンションと呼ばれる手法を用いた研究 [6, 7] が注目されている。アテンションとは、深層学習において分類・予測を行う際、出力に直接結びつく入力を探る手法で、このアテンションにより、出力に貢献する特徴は何かを視覚的にわかりやすくなっている。最新の研究では、アテンション計算を層ごとに行い、より分類・予測精度を高めた研究 [8] や、アテンションのみで構築された深層学習 [9] なども登場している。しかし、アテンションはあくまで入力と出力の関係のみに注目し、内部でどのような学習が行われているかは考慮していない。

そこで、自身の研究 [10] では、テキストベースの深層学習について、層ごとの学習の流れを単語情報として表し、人間が理解できる形に直すことで、分類基準の理解のための、学習ネットワークの解釈を支援するシステムの開発を目的とし、一定の成果を得ることができた。一方で、この時使用した深層学習が、構造は単純だが特徴量（テキスト中の単語）の有無だけ学習

し、単語の出現の時系列や順序を一切考慮しないものだったため、学習ネットワークの解釈が一定までしか得られなかった。

したがって、本研究ではこのような問題意識の下、文章（テキスト）の分類問題を例として、時系列関係を含めた分類に寄与する出力ごとの特徴を抽出できるように、再帰的ニューラルネットワークを用いるようにした。また、深層学習が持つ、ネットワークの各層ノード間の関係にマルコフ性が存在する [11] 性質を利用して、学習ネットワークを HMM に当てはめることで、ネットワークの各層に付けられた重みの値から、タイムステップごとの入力単語の順序列に対する尤度を算出、そこから、尤度の高い単語の順序列を分類のパターン、つまり出力を導くルールとして抽出するシステムの構築を目指す。また、システムでは抽出された分類パターンを可視化するインターフェースを備えている。

3 HMMを利用した再帰的深層学習ネットワークからの分類パターン抽出・可視化システム

本章では、本研究で開発した HMM を利用した深層学習ネットワークからの分類パターン抽出・可視化システム（以後、提案システム）について、システムの構成とその詳細について述べる。

3.1 提案システムの構成

提案システムでは、まず、図1に示すように、各分類先ごとにラベル付けしたテキスト集合を LSTM にて分類し、その分類先を導いた学習ネットワークを HMM に当てはめ、提案システムの分類パターンの抽出処理部によって各出力（分類先）を導く分類パターンの尤度に基づく抽出を行う。最後に、システムの利用者は、システムの可視化処理部によって得られた学習ネットワークの表示を自分が見やすいように調整し、分類パターンを可視化する。また、システムでは分類パターンの意味を理解しやすくするための機能（解釈支援機能）を利用できる。

3.2 深層学習による学習ネットワークの形成

3.2.1 テキスト中の単語のベクトル化

深層学習で学習を行う前に、テキストデータはテキスト中の単語を取り出したあと、単語を One hot 法 [12] と呼ばれる手法に従い単語ベクトルの羅列に直す。そして、テキスト中の各単語をその単語ベクトルに置き換え、深層学習への入力データとする。

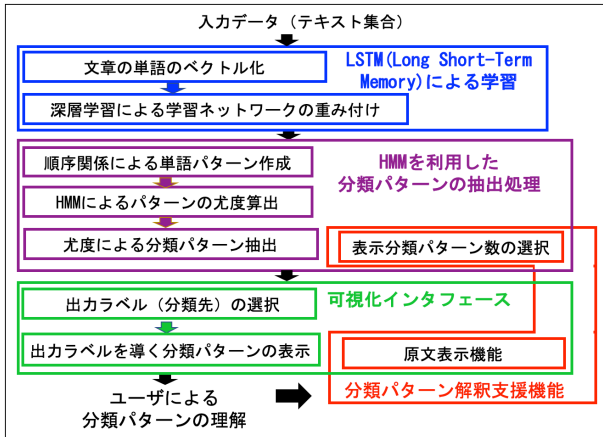


図 1: 提案システムの構成

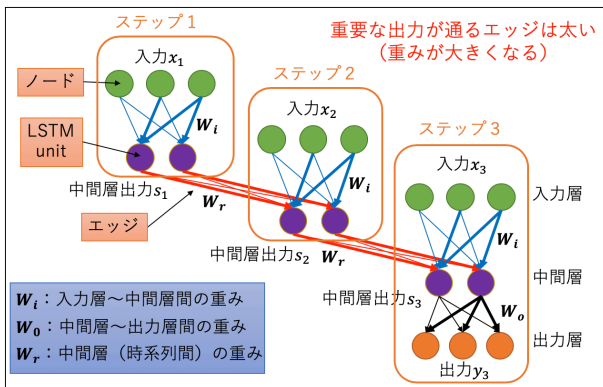


図 2: 再帰的深層学習の学習ネットワークと学習の様子

3.2.2 学習によるネットワークの重み付け

One hot 法によって単語ベクトルの羅列に変換され、分類先ごとにラベル付けされたテキストデータは、LSTM でそれぞれの出力層ノード（分類先）を導くネットワークへの重み付けがされていく。その様子を図 2 に示す。入力文章は各単語がベクトル化され、タイムステップごとに単語ベクトルが順番に入力されていく。また、LSTM での分類時は、最後の単語が入力されたタイミングで、出力層から分類結果が出力される。

3.3 HMM を用いた学習ネットワークからの分類パターンの抽出・可視化処理

3.3.1 LSTM の HMM への変換

提案システムの分類パターンの抽出処理では、LSTM によって得られた学習ネットワークを図 3 のように、ひとつの HMM として処理を行う。

まず、分類パターンの候補として、LSTM への入力に使用した全単語の組み合わせを作成する。この時、組み

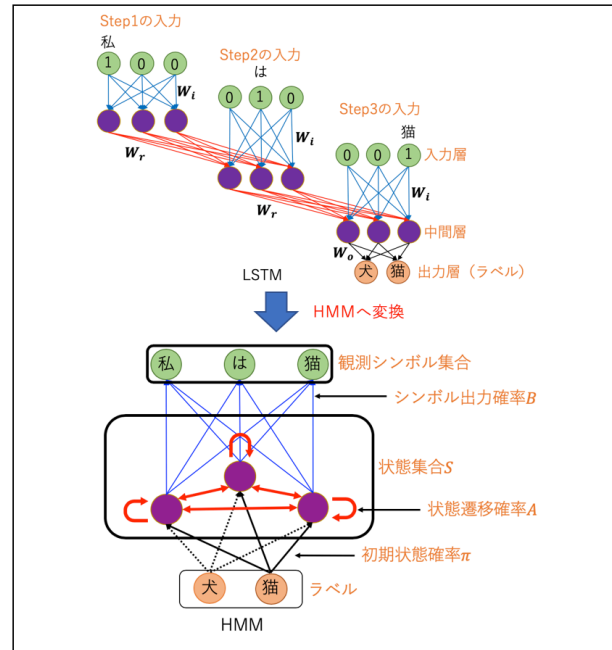


図 3: LSTM の HMM への変換

合わせの条件として以下を満たす単語列を候補とする。

- 分類パターン候補の長さ（単語数）は任意で決めた長さで揃えるとする
- 分類パターン候補の単語の順序は実際のテキスト中の単語の出現順序に基づくものとする

次に LSTM の入力層ノードを HMM の観測シンボル集合、中間層ノード（LSTM ユニット）を状態集合 $S = \{s\}$ とし、同様に中間層の（再帰的処理による）時系列間の重みを状態遷移確率 A 、入力層中間層間の重みをシンボル出力確率 B とする。そして、中間層出力層間の重みを初期状態確率 π とするが、この π はその時選択するラベル（分類先）によって変わる。この時、観測シンボルによる観測系列（前述した分類パターン候補）を $O = o_1, o_2, \dots, o_T$ (T は観測系列の長さ（前述した分類パターン候補の長さ）)、状態数（中間層ノード数）を N （状態番号は i, j ）と置くと、状態遷移確率 A は式 (1)、シンボル出力確率 B は式 (2)、初期状態確率 π は式 (3) となる。

$$A = \{a_{ij} | a_{ij} = P(s_{t+1} = j | s_t = i)\} (1 \leq i, j \leq N) \quad (1)$$

$$B = \{b_{ij}(o_t) | b_{ij}(o_t) = P(o_t | s_{t-1} = i, s_t = j)\} (1 \leq i, j \leq N, 1 \leq t \leq T) \quad (2)$$

$$\pi = \{\pi_i | \pi_i = P(s_0 = i)\} (1 \leq i, j \leq N) \quad (3)$$

この時、 A, B, π で構成される HMM を式 (4) のように略記する。

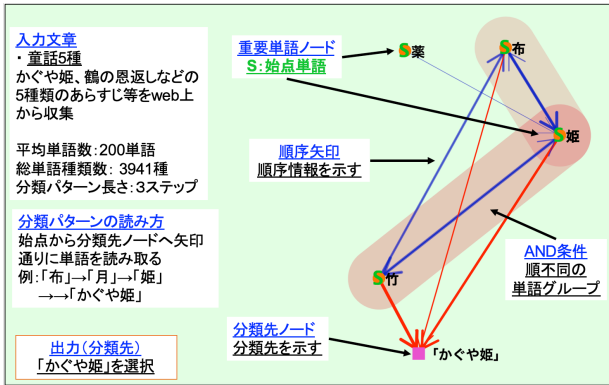


図 4: 提案システムの画面例

$$\lambda = \{\pi, A, B\} \quad (4)$$

最後に、分類パターン候補の尤度を算出し、尤度の高い順に分類パターンとして抽出する。あるラベル a に対して、分類パターン候補 $O = o_1, o_2, \dots, o_T$ がある時、分類パターン候補 O の尤度 $P(O|\lambda)$ は、式 (5) で算出される。

$$\begin{aligned}
 P(O|\lambda) &= \sum_{all S} P(O|\lambda)P(O|S, \lambda) \\
 &= \sum_{all s_0 \dots s_T} \pi_{s_0} a_{s_0 s_1} b_{s_0 s_1}(o_1) \cdot a_{s_1 s_2} b_{s_1 s_2}(o_2) \cdot \\
 &\quad \dots \cdot a_{s_{T-1} s_T} b_{s_{T-1} s_T}(o_T) \quad (5)
 \end{aligned}$$

こうして LSTM の学習ネットワークから、時系列を考慮した単語の並びとしての分類パターンを抽出し、各分類先への寄与の強さを尤度で表すことができる。なお、式 (5) の計算量は $O(2TN^T)$ であるため、深層学習モデルが大きくなると計算量が非常に大きくなってしまふ。そのため、本来は forward-backward アルゴリズム等を用いるが、今回使用したモデルは比較的規模が小さいため、式 (5) で計算を行った。

3.3.2 分類先を導く分類パターンの可視化

提案システムの可視化処理部では、分類先に強く結びつく、尤度の高い分類パターン集合が表示される。例として、5種類の童話のあらすじに関するテキスト集合の分類を行った場合の、提案システムのメイン画面を図 4 に示す。表示分類先は「かぐや姫」を選択している。図 4 では、分類パターン中の単語の流れを矢印の向きで表し、分類パターンを構成する単語をノードで表している。また、尤度の大きさを矢印の太さで表している。

表 1: 学習ネットワーク表示機能

機能名	効果
表示分類パターン数の増減	一つの分類先につき、いくつの分類パターンを表示するかを選択する
分類パターンの長さの選択	分類パターン候補作成時に、分類パターン候補の長さを選択する
原文表示機能	分類パターン（順序を考慮した単語組み合わせ）が、原文中でどのように出現しているかを表示する

表 2: テキストデータの詳細

データ名	内容
人工データ (6パターン)	3種類の重要記号「A0」「A1」「A2」を1つずつ使った組み合わせ6パターン(例「A0」→「A1」→「A2」, 「A1」→「A2」→「A0」など)をラベルとし、重要記号間にランダムで選択したノイズ記号「N0」から「N999」を5つずつ挿入したものを1テキストとして、パターンごとに5,000(計30,000)テキスト用意した
人工データ (24パターン)	4種類の重要記号「A0」「A1」「A2」「A3」を1つずつ使った組み合わせ24パターン(例「A0」→「A1」→「A2」→「A3」, 「A1」→「A2」→「A3」→「A0」など)をラベルとし、重要記号間にランダムで選択したノイズ記号「N0」から「N999」を5つずつ挿入したものを1テキストとして、パターンごとに1,000(計24,000)テキスト用意した

3.4 分類パターン解釈支援機能

システムには、利用者が抽出された分類パターンの解釈しやすいように、その表示内容を調整できる機能がある。その主なものを表 1 に示す。

4 分類パターン抽出システムの有効性の検証実験

本章では、提案システムについて、分類パターンの抽出に焦点を当て、抽出された分類パターンが、テキストデータの理解（学習ネットワークの解釈を行う）を目的とした上で、各分類先に特徴的な分類パターンが抽出できているかを検証した実験について述べる。なお、本実験では、分類パターンを「順序を考慮した任意の長さの異なる単語の組み合わせ」と定義する。

表 3: 学習 LSTM モデルの詳細

データ名	抽出した 単語集合	入力層 ノード 数	中間層 LSTM ユニッ ト数	出力層 ノード 数
人工デー タ (6 パ ターン)	1,003 記号	1,003	10	6
人工デー タ (24 パ ターン)	1,004 記号	1,004	10	6

表 4: 人工データ (6 パターン) の偏差値

ラベル	尤度 10^{-3}	平均 10^{-3}	偏差値	標準偏差 10^{-3}
A0 → A1 → A2	4.56	0.10	149.4	0.45
A0 → A2 → A1	5.20	0.10	149.5	0.51
A1 → A0 → A2	4.75	0.10	149.4	0.47
A1 → A2 → A0	5.39	0.12	149.4	0.53
A2 → A0 → A1	5.06	0.11	149.4	0.50
A2 → A1 → A0	4.70	0.10	149.4	0.46
平均	4.94	0.11	149.4	0.49

4.1 実験準備

4.1.1 使用テキストデータと学習モデル

分類パターン抽出の対象として、決まったパターンを持つ記号の組み合わせで構成された人工データを 2 つ用いる。データの詳細について表 2 に示す。

続いて、深層学習として使用した LSTM モデルの概要を表 3 に示す。また、抽出する分類パターンは長さ 3 で個数はラベル (分類先) ごとに 100 とする。なお、2 つの人工データに対する LSTM の分類精度はどちらも 100% である。

4.2 実験手順

実験は著者 1 名で行い、テキスト「人工データ (6 パターン)」、「人工データ (24 パターン)」について、提案システムの分類パターン抽出処理部によって、各分類先ごとに尤度の高い順に 100 個分類パターンを抽出した。そして抽出された分類パターンの尤度の偏差値や標準偏差を算出し、決められたパターンを持つ分類パターン (各ラベルごとに決めたパターンのうちノイズを含まないもの) とその他のノイズを含む分類パターンの偏差値等を比較した。

なお、分類パターンの長さは 3 なので、テキスト「人工データ (6 パターン)」では決められたパターンを持つ分類パターンは各パターンに 1 つずつ (決められたパターンの長さが 3 のため) となり、テキスト「人工データ (24 パターン)」では、決められたパターンを持つ分類パターンは各パターンに 4 つずつ (決められたパターンの長さが 4 のため) となる。

4.3 結果と考察

テキスト「人工データ (6 パターン)」について抽出された分類パターンの尤度上位 100 の偏差値を出し、そのうち、各ラベルごとに決められたパターンを持つ分類パターンの偏差値、100 個中の尤度の平均、標準偏差を表 4 に示す。また、テキスト「人工データ (24 パ

ターン)」について抽出された分類パターンの尤度上位 100 の偏差値を出し、そのうち、各ラベルごとに決められたパターンを持つ分類パターンの偏差値、100 個中の尤度の平均、標準偏差を表 5 に示す。ただし、「人工データ (24 パターン)」の方は決められたパターンを表す分類パターンが各ラベルごとに 4 つずつあるため、偏差値はその 4 つの平均とする。

表 4、表 5 より、決められたパターンを持つ分類パターンは、その他のノイズを含む分類パターンより尤度や偏差値が大きくなることがわかった。これは、提案システムでは、テキスト特有の特徴を表す、学習ネットワークの解釈に適した分類パターン (人工データにおける決められたパターンを持つ分類パターン) が、その他の学習ネットワークの解釈に適さない分類パターン (ノイズ等を含む、不必要な情報を持った分類パターン) より明確な差を持って抽出できることを示す。これにより、抽出する分類パターンの尤度に閾値を設けて可視化すれば、提案システムの利用者は、テキストの分類に明確に寄与する分類パターンのみを対象とした解釈ができる。

また、表 4 と表 5 を比較すると、表 4 では、決められたパターンを持つ分類パターンの尤度が平均の 10 倍前後になっており、偏差値も 150 近い値となっている一方、表 5 では、決められたパターンを持つ分類パターン尤度が平均の 2 倍から 3 倍程で、偏差値も 100 を下回っている。これは、表 4 の決められたパターンを持つ分類パターンは、完全にその分類先にしか存在していないので尤度が高くなるが、表 5 の決められたパターンを持つ分類パターン 4 つは、一つひとつは他の分類先にも存在しているため、表 4 より尤度が高くならなかったからと思われる。よって、学習させるテキストの長さに適した長さの分類パターンを選ぶ方が、より尤度の高い特徴的な分類パターンを抽出できると考えられる。

5 おわりに

本研究では、複数のテキストデータの分類を、単語の順序関係を学習できる深層学習である LSTM で行い、

表 5: 人工データ (24 パターン) の偏差値

ラベル	尤度 10^{-7}	平均 10^{-7}	偏差値	標準偏差 10^{-7}
A0 → A1 → A2 → A3	4.76	1.66	93.1	0.72
A0 → A1 → A3 → A2	4.93	1.84	93.8	0.71
A0 → A2 → A1 → A3	5.36	1.96	91.2	0.82
A0 → A2 → A3 → A1	4.71	1.90	88.0	0.74
A0 → A3 → A1 → A2	5.28	1.92	93.7	0.77
A0 → A3 → A2 → A1	4.32	1.66	90.6	0.66
A1 → A0 → A2 → A3	5.41	1.78	93.4	0.84
A1 → A0 → A3 → A2	5.21	1.77	94.0	0.78
A1 → A2 → A0 → A3	5.32	1.74	93.5	0.82
A1 → A2 → A3 → A0	5.65	2.02	92.7	0.85
A1 → A3 → A0 → A2	4.73	1.67	93.5	0.70
A1 → A3 → A2 → A0	4.45	1.54	93.3	0.67
A2 → A0 → A1 → A3	5.45	1.92	92.2	0.84
A2 → A0 → A3 → A1	4.70	1.79	91.0	0.71
A2 → A1 → A0 → A3	5.58	2.06	92.7	0.82
A2 → A1 → A3 → A0	4.55	1.79	92.5	0.65
A2 → A3 → A0 → A1	5.53	1.80	93.7	0.85
A2 → A3 → A1 → A0	5.55	2.12	92.1	0.81
A3 → A0 → A1 → A2	5.32	1.79	94.3	0.80
A3 → A0 → A2 → A1	3.83	1.44	91.2	0.58
A3 → A1 → A0 → A2	5.93	2.10	93.6	0.88
A3 → A1 → A2 → A0	5.24	1.83	92.5	0.80
A3 → A2 → A0 → A1	5.54	1.75	93.4	0.87
A3 → A2 → A1 → A0	4.92	1.72	94.4	0.72
平均	5.13	1.77	93.2	0.77

学習ネットワークの解釈を行うための、分類パターンの抽出システムの構築を目的とした。本研究の特徴として、複雑な再帰的深層学習のネットワーク構造をHMMに当てはめて処理することで、容易に、学習された特徴量の順序情報を抽出できる点が挙げられる。提案システムの有効性を確かめる検証実験では、提案システムで抽出された分類パターンの尤度について、決められたパターンを表す分類パターンとノイズを含む解釈に向かない分類パターンを比較し、決められたパターンを表す分類パターンの方が、尤度が特徴的に大きく抽出できていることから、提案システムで抽出される分類パターンはテキストの特有の特徴を有しているものが表示されやすいと結論づけた。今後の研究では、実際に抽出された特徴を用いて、それらからどのような学習ネットワークの解釈が行えるかを重視して進めていく予定である。

参考文献

[1] ボレガラ ダヌシカ, “自然言語処理のための深層学習”, 人工知能学会誌, Vol.29, No.2, pp.195-201, 2014

[2] Ebru Arisoy, Tare N. Sainath, Brian Kingsbury, Bhuvaba Ramabhadran, “Deep Neural Network Language Models”, In Proceedings of the NAA-CLHLT Workshop, Will We Ever Really Replace the N-gram Model?, pp.20-28, 2012

[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition”, In Proceedings of the IEEE, 1998

[4] Matthew D. Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks”, In Proceedings of ECCV ’14, pp.818-833, 2014

[5] 西銘 大喜, “ディープニューラルネットワークによる画像からの表情表現の学習”, 第29回人工知能学会全国大会, 3L4-3, 2015

[6] M Daniluk, T Rocktaschel, J Welbl, S Riedel, “Frustratingly Short Attention Spans in Neural Language”, ICLR, 2017

[7] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need”, CoRR, vol. abs/1706.03762, 2017

[8] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. “Convolutional sequence to sequence learning”, arXiv preprint arXiv:1705.03122v2, 2017

[9] A.Vaswani, N.Shazeer, N.Parmar, J.Uszkoreit, L.Jones, A.N.Gomez, L.Kaiser, I.Polosukhin, “Attention Is All You Need”, In the Annual Conference on Neural Information Processing Systems (NIPS), 2017

[10] 安藤雅行, 河原吉伸, 砂山渡, 畑中裕司, “テキストベースの深層学習における分類パターンの解釈支援”, 知能と情報(日本知能情報フェジィ学会誌), Vol.31, No.4, pp.779-787, 2019

[11] R. Shwartz-Ziv and N. Tishby. “Opening the black box of deep neural networks via information”, arXiv preprint arXiv:1703.00810, 2017

[12] M. Feurer, A. Klein, K. Eggenberger, J. Springenberg, M. Blum, and F. Hutter. “Efficient and robust automated machine learning”, In Neural Information Processing Systems (NIPS), 2015

専門知識がないユーザのための対話的政治情報システムの提案

Proposal of a Interactive Political Information System for Non-specialist

渋谷 英潔¹ 石下 円香¹ 神門 典子^{1,2}
Hideyuki Sibuki¹ Madoka Ishioroshi¹ Noriko Kando^{1,2}

¹ 国立情報学研究所 ² 総合研究大学院大学
¹National Institute of Informatics ²SOKENDAI

Abstract: To make non-professional users aware of argument viewpoints, we aim to develop a multi-faceted exploratory search system on political information, which involves a facet based on argument reasons. As a fundamental study, we considered support functions the system should have, and researched keywords using Japanese local assembly minutes. Using grammatical templates that are combinations of particles, suffixes and auxiliary verbs, phrases that seem to describe reasons/purposes of political arguments were extracted from the minutes. We collected noun sequences in the phrases as keywords, and the type number of keywords was 156,697 in total.

1 はじめに

2017年のユーキャン新語・流語賞において「フェイクニュース」がTOP10に選ばれたのは記憶に新しい。我々はこれまで情報信憑性判断支援に関する研究 [1, 2, 7] を行い、NTCIR QA Lab PoliInfo[3] では質問応答や自動要約などの技術を用いてフェイクニュースに対処するタスクを開催した。これらの取り組みから、ニュースなどの信憑性を判断するためには一次情報、例えば、議員の発言であれば議会会議録などにあたる必要があることが分かっている。しかしながら、議会会議録は非常に長く、信憑性判断の度に全てを読むのはコストが大きい。そのため、必要とする情報に容易にアクセスできる技術が重要である。

東京都をはじめ多くの地方自治体がインターネット上で検索システムとともに議会会議録を公開している。図1は東京都議会の会議録検索システムであり、利用者に検索クエリを入力させる一般的な検索システムである。複雑な情報要求に対するクエリ入力型の情報検索において大抵の場合、利用者が望む情報に1回の検索で辿りつくことは少なく、望む情報に辿りつくまで幾つかの検索クエリを試して対話的に絞り込んでいくことが知られている。しかしながら、政治に詳しくない利用者にとっては、どのような語句を検索クエリとして用いればよいのか分からず、場合によっては最初



図 1: 東京都議会の会議録検索システム

の検索クエリすら思いつかないという問題がある。また、図1のシステムには発言者(議員)や会議名などで絞り込む機能も備わっているが、やはり政治に詳しくない利用者にとっては活用することが困難である。

こういった問題に対処するため、我々は、専門知識がない利用者が対話的に適切な情報にアクセスできる政治情報システムを目指している。本稿では、我々が目的とする対話的政治情報システムの基礎的検討として、システムに必要な機能を考察し、地方議会会議録を用いて調査した結果を報告する。

2 関連研究

信憑性判断支援や意思決定支援に関する従来研究には以下のものがある。情報分析システム WISDOM X[4]は、ファクトイド型、なぜ型、どうなる型、定義型の4種類のQAシステムで構成されており、システムを切り替えることで検索クエリとの関係性を考慮した結果を提示した。伊木ら [5]は、レビュースパムに対処するため、類似性、協調性、集中性、情報性という4つのスコアを提示することで、レビューの信頼判断支援を行った。佐藤ら [6]は、国会会議録を意思決定支援に活用することを目的としてディベートをするシステムを開発した。松本ら [7]は、着目言明に関連する条件と結論を抽出し対話的にマップを広げていくことで情報信憑性判断を支援する方法を提案した。大塚ら [8]は、コミュニティQAを用いてクエリ拡張することで利用者が想起できないクエリを支援した。

また、議会議録を用いた分析には、以下の研究がある。小柴ら [9]は、政策立案プロセスにおける議会と行政の相互作用を分析し、大南ら [10]、国務大臣としての資質に欠ける人物の特徴を見出すことを目的として議会議録を分析した。関 [11]は世代による政治ニュース記事の閲覧傾向の違いを分析し、中川ら [12]は、ランダムフォレストを用いて議会議録のイデオロギーを分析した。しかしながら、これらの研究は我々の目的と異なる。

3 専門知識がない利用者にとっての問題点

1節で述べたように、専門知識がない利用者にとって適切な検索クエリを想起することは容易ではない。そして、仮に望む情報を含む文書が検索されたとしても、それが望む情報を含んでいるか判断できるとは限らない。例えば、議会議録には「第百三十三号議案に反対する立場から討論を行います」といった記述があるが、「第百三十三号議案」がどのような内容であるのか分からなければ、ここでの議論が、望む情報と関連しているか判断できない。また、例えば「築地市場移転」と「水銀検出」という語句がシステムから提示されたとしても、「築地市場で水銀が検出されたから移転すべき」なのか「築地市場の移転先で水銀が検出されたから移転中止すべき」なのか分からない。さらに、「リゾート施設の建設に50億円のコストがかかるので反対であ

る」という根拠をもとにした意見が示されても、相場が分からなければ判断ができないという問題もある。

したがって、専門知識がない利用者にとっての問題点は大きく以下の4つが存在する。

- (A) 適切な検索クエリが想起できない
- (B) 提示された語句の定義が分からない
- (C) 提示された語句間関係性が分からない
- (D) 提示された根拠の妥当性を判断できない

まず、問題点(A)については、システムからキーワードを提示するファセット検索が有効であると考えられる。問題点(B)については、定義をオンマウスでのポップアップなどで表示する仕組みが考えられる。問題点(C)については、信憑性判断支援の観点から、原因と結果、目的と根拠といった関係性が重要である。しかしながら、一般的な議論において、目的(結果)が比較的容易に推測できる場合、根拠(原因)のみを述べることが多い。例えば、「リゾート施設を建設することで観光客の増加が見込める」と述べた場合、「建設に賛成である」という話者の意見が容易に推測できる。ここで、なぜその推測が可能かを内省すると、「観光客が増加することで、その地域で使うお金が増加し収益につながる」という考えに至る。また、「観光客が増加することで治安悪化の恐れがある」が反論であると推測できるのは、「治安が悪化することで、その地域の活動が困難になり収益が減少する」という考えが基本にあるからだと考えられる。そこで、議会における目的は金銭的収支により近似できるという仮説¹を立てる。議会における金銭的収支は、予算報告や決算報告という形式で公開されていることから、それらと対応付けることで指標の一つとすることができる。仮に、議論における全ての根拠に対して金銭的収支による目的を提示することができれば利用者の判断支援につながると考えられる。最後、問題点(D)については、問題点(C)で述べた金銭的収支による提示を実現することで、類似事例から相場を提示できると考える。

次節以降、問題点(A)に関連して議員を特徴づけるキーワード、問題点(C)に関連して目的と根拠のキーワードを調査する。

¹もちろん、目的の全てが金銭的収支につながるわけではないが、政治における議題の多くが金銭的収支に関係すると考えられる。

表 1: 抽出された議員ごとのキーワード

M 字カーブ現象 ICT 関連のサービス ICT 端末の導入	MICE の可能性 売店の全店リニューアル 大阪城公園のパークマネジメント	幼児教育のノウハウ 子育て安心プラン 六次産業化
JK ビジネス被害 HIV の感染リスク 新規 HIV 感染者	府中の小児総合医療センター 二カ月間 東京農業振興プラン	和式のトイレ 都民ファーストの視点 三カ所設置
予定の羽田空港アクセス線 民間のワーキングマザー 保護者のニーズ	CO2 の削減 CO2 の吸収 旅行者のニーズ	U20 加盟都市 AED の知識 AED の設置
日本のエンジン 都内鉄道駅のエレベーター 新規ビジネスの育成	武蔵野の森スポーツ施設 浸透型トレンチ管 三百トン	性的マイノリティーの方々 手話のアルファベット 自分のスペシャルニーズ
JR 南武線の矢川 JR 西国立駅周辺 二百十ヘクタール	東京のプレゼンス 簡易トイレの組み立て 環境対策のモデル	TooGoodToGo の紹介 容器のリサイクル対策 民間のレストラン
保育園の ICT システム 3D テレビ 東京都 ICT 戦略	民間のツアー 東京港の運河エリア 都民ファーストの代表質問	SNS の活用 JR 信濃町駅 入院サポートセンター
方のイロハ 地域コミュニティの重要性 太陽光パネルの設置	知事の COPD 区の AED ATM の横	SDGs のシナジーの強化 SDGs の取り組み IOC の方針のもと
医療機関間の ICT 臨床データの蓄積 難病医療センター	都市計画マスタープラン等 教育委員会のホームページ さきのオリンピック	両者のバランス 東京マンハッタン計画 東京ベイエリアビジョン策定
PFI 法の改正 二〇二〇パラリンピック 知事のリーダーシップそのもの	公共施設の ZEB 化 略称サ高住 防犯ブザーの配布	目黒のケース 福祉のサービス 大会のボランティア活動
六カ月の女の子 保育サービスの違い 馬込三寸ニンジン	光のページェント LED 陸上競技のトップアスリート 算数ドリルの周知	八万ヘクタールの森林 森林管理システム 十メートルの津波
練馬駅ハビリテーション病院 保健医療サービス 非常勤の巡回指導チーム	公衆無線 LAN 多摩地域の道路ネットワーク 多摩メディカルキャンパスの存在	通り一遍のアンケート 住民サービスの原則 逆流防止の受水タンク
八キロメートル区間 排水ポンプ整備運用計画 二百十二平方キロメートル	米国大統領ケネディ 二千ヘクタール 東京のプロモーション	二番目のダブル A 最上級のトリプル A 東京版 EMP
PDCA のサイクル 体のバランス 数時間のプログラム	JR の中央線 多摩都市モノレール推進室 障害者雇用エクセレントカンパニー賞	方改革プランづくり 保険外サービス 日本語教育のスキル
最新の ICT 情報共有システムの運用 社会的養育ビジョン	G7 サミット JR 中央線 JR 横浜線	日の出山エリア 東京の自然公園ビジョン 都のパブリックコメント
東京 DMAT カー イギリスの EU 脱退 日 EU 経済連携協定	防災事業のスピードアップ 訪問看護ステーション 小規模事業所アンケート調査	ICT 環境整備のあり方 保育園のエリア 二十キロメートル圏内
二年三カ月間 調べ学習のコンクール 授業実践モデル校	MUD の考え方 BRT の魅力 BRT の停留施設	臨海部都市開発プロジェクト 旅行者の行動パターン 日本の最先端テクノロジー
アメリカのスペース X 社 未来の産業のリーダー 飛躍のチャンス	放課後子ども総合プラン 認知症サポーターの活用 認知症サポーター	GHQ の統治 CO2 の削減効果 連合赤軍リンチ殺人事件
水害版 BCP の作成 BCP の重要性 東京女子医科大学東医療センター	中休みのマラソン 全国学力テスト 全員対象の学力テスト	RE 一〇〇 六千キロワット 民間のアイデア
二つの SNS の相談 児童虐待の LINE 相談 子供たちのニーズ	GAP の認証取得 GAP の取得 GAP の意義	LP ガス空調機 八王子南バイパスの整備 道路ネットワークの整備効果
代替フロン HFC 廃棄のサイクル 二十点のレベル	PDCA サイクルの徹底 TOEFL スコアランキング 百九十億ドル	東京の集積のメリット 船着き場の案内サイン 四十三キロ

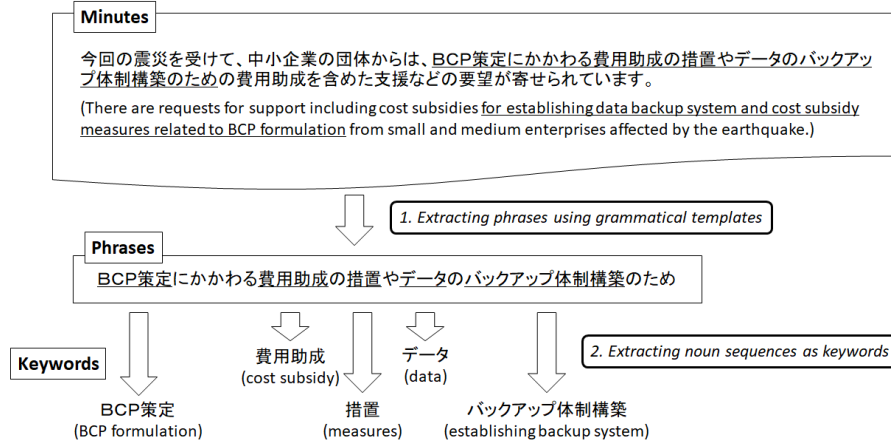


図 2: 目的や根拠の手がかり表現からキーワードを抽出する流れ

4 地方議会会議録の予備調査

4.1 議員を特徴づけるキーワード

地方議会会議録コーパスプロジェクト²において都道府県議会会議録検索システム「ぎ～みる」が公開されており、TF-IDF を用いて議員ごとの特徴語を抽出している。これらの特徴語には役立つものが多いが、中には「ゴルフ」、「一般病院」、「口座振替」といった、やや漠然としていて、もう少し補足説明がほしいものも存在する。そこで、「名詞句の名詞句」という形式のキーワードを許すことで、より適切なキーワードが抽出できないか検討する。

まず、質疑における発言には発言者の考えが強く反映されていると仮定した。質疑には代表質問と一般質問があるが、代表質問は所属する会派を代表して質問するものであり、述べられている内容が必ずしも質問者自身の考えでない可能性がある。それゆえ、一般質問における発言を対象に以下の手順で抽出を行った。最初に、MeCab³を用いて形態素解析を行い、品詞が「名詞」で、品詞細分類1が「代名詞」、「副詞可能」、「非自立」以外の連続する形態素列を名詞句として抽出した。次に、品詞が「助詞」で品詞細分類1が「連体化」の形態素の前後にある名詞句列を1つの名詞句としてまとめた。1つの質問を1つの文書とみなして、TF-IDFを以下の式で計算した。

$$tfidf(i, j) = tf(i, j) \cdot idf(i) \quad (1)$$

$$tf(i, j) = \frac{n(i, j)}{\sum_k n(k, j)} \quad (2)$$

²<http://local-politics.jp/>

³<https://taku910.github.io/mecab/>

$$idf(i) = \log \frac{|D|}{|d: d \ni t_i|} \quad (3)$$

$n(i, j)$ は文書 d_j における単語 t_i の出現回数、 $|D|$ は総文書数、 $|d: d \ni t_i|$ は単語 t_i を含む文書数である。

一般にひらがなのみの文字列よりも漢字、カタカナ、アルファベットを含む文字列の方がキーワードになりやすいため、名詞句に含まれる字種ごとの重みを全て足し合わせた値を名詞句の重み $w(i)$ とした。

$$w(i) = 1 + w_C \cdot C(i) + w_K \cdot K(i) + w_A \cdot A(i) \quad (4)$$

w_C, w_K, w_A はそれぞれ漢字、カタカナ、アルファベットが含まれる場合の重みで、 $C(i), K(i), A(i)$ はそれぞれ漢字、カタカナ、アルファベットが名詞句中に含まれる場合は1、そうでなければ0を返す関数である。

キーワードの長さは長すぎても短すぎても不適切であると考え、名詞句中の形態素数が α 以上 β 以下のものを適切とした。 α より不足または β より超過している形態素数を $e(i)$ として、最終的なスコアを以下の式で求めた。

$$sc(i, J) = \frac{\sqrt{w(i)}}{e(i)} tfidf(i, j) \quad (5)$$

平成30年第1回定例会議会会議録を対象に調査を行った。一般質問をした議員は60名であった。暫定的に、 $w_C = 1, w_K = 2, w_A = 3, \alpha = 3, \beta = 5$ とした時の結果を表1に示す。表中の1つの欄が1人の議員に対応し、各議員の上位3件のキーワードを記している。「幼児教育のノウハウ」や「HIVの感染リスク」など、より明確なキーワードが抽出できている。しかしながら、現状では定量的な評価は難しく、今後、対話的政治情報システムに実装してシステムを通して有効性を評価したいと考えている。

表 2: 根拠と目的のテンプレート一覧

根拠	
-から	V/A (基本形) + から
-ので	な + ので
-ため	V/A (基本形) + ので
	A (基本形) + ため
	た + ため
	な + ため
-おかげで	の + ため
	V/A (基本形) + おかげ + で
	な + おかげ + で
	の + おかげ + で
-せいで	V/A (基本形) + せい + で
	な + せい + で
	の + せい + で
-ばかりに	た + ばかり + に
	な + ばかり + に
目的	
-ため	V (基本形) + ため
-よう	V (基本形) + よう
-べく	V (基本形) + げく
-と	う + と
	まい + と

4.2 目的と根拠のキーワード

一般に、目的や根拠を示す手がかりは「から」や「ため」といった文(節)末の機能的表現により示される。しかしながら、地方議会では方言など地域ごとの言い回しが用いられる場合があり、必ずしも一般的な表現が用いられているか自明ではない。そこで、地方議会会議録コーパスを用いて、平成23年4月の統一地方選挙から平成27年4月の統一地方選挙の前(平成27年3月)までの4年間における全国47都道府県議会の本会議を対象に、一般的な機能的表現(テンプレート)によりキーワードが抽出できるか調査を行った。

図2に、目的と根拠のテンプレートを用いたキーワード抽出の流れを示す。「から」や「ため」といった根拠や目的を表すテンプレートを用いて、根拠や目的を含む節を特定し、特定された節内の名詞連続をキーワードとして抽出する。MeCab⁴を用いて形態素解析を行い、文頭または句読点を手がかりに節の境界を決定した。根拠と目的を表すテンプレートは、文献[13]を参考として、表2に示した根拠15種類、目的5種類の計20種類を用いた。

テンプレートにより抽出された節の異なり数を表3に都道府県ごとに示す。議会会議録において現れるテンプレートの傾向は異なり、「おかげで」や「せいで」のような主観を含むテンプレートの頻度は少なかった。

⁴<https://taku910.github.io/mecab/>

しかしながら、都道府県による差異は小さく、共通のテンプレートを用いて抽出できることを確認した。

図3と図4に、根拠と目的のそれぞれのテンプレートで抽出された節の回数上位10件を示す。また、図5に、根拠と目的のテンプレートにより抽出されたキーワードを文字長別に上位10件ずつ示す。年度が異なるため表1のキーワードと直接比較はできないが、「名詞句の名詞句から」や「名詞句の名詞句のため」といった表現はあまり見受けられず、議員ごとのキーワードと対応付けるには何らかの工夫が必要であると思われる。

5 まとめ

本稿では、専門的な知識を持たない利用者が求める情報にアクセスする際に生じる問題点について述べ、それを支援するための機能について考察した。また、地方議会会議録を用いて抽出した、議員を特徴づけるキーワードと、目的と根拠のキーワードを報告した。今後、考察した機能やキーワードを実装した対話的政治情報システムを開発し、その有効性を実験する予定である。

謝辞

本研究は科研費16H01756の助成を受けたものです。また、地方議会会議録コーパスを提供していただいた小樽商科大学の木村泰知准教授に感謝します。

参考文献

- [1] 渡木英潔, 永井隆広, 中野正寛, 石下円香, 松本拓也, 森辰則. 情報信憑性判断支援のための対話型調停要約生成手法. 自然言語処理, Vol. 20, No. 2, pp. 75-104, (2013).
- [2] 渡木英潔, 中野正寛, 宮崎林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則. 情報信憑性判断支援のための Web 文書向け要約生成タスクにおけるアノテーション. 自然言語処理, Vol. 21, No. 2, pp. 157-212, (2014).
- [3] Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ootake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando, Tatsunori Mori, Harumichi Yuasa, Satoshi Sekine, and Kentaro Inui. Overview of the NTCIR-14 QA Lab-PoliInfo Task. In Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies, pp. 121-140, (2019).
- [4] 水野淳太, 田仲正弘, 大竹清敬, 呉鍾勲, Julien Kloetzer, 橋本力, 鳥澤健太郎. 大規模情報分析システム WISDOM X, DISAANA, D-SUMM. 言語処理学会第23回年次大会発表論文集, pp. 1077-1080, (2017).

表 3: テンプレートにより抽出された都道府県ごとの節の数

都道府県	発言数	-から	-ので	-ため	-おかげで	-せいで	-ばかりに	-ため	-よう	-べく	-と
北海道	108,344	566	504	1,641	1	0	0	5,935	4,755	36	1,958
青森	96,095	1,088	2,163	2,324	3	1	0	5,839	3,625	74	610
岩手	83,436	825	2,343	1,900	4	1	0	3,616	3,669	56	2,153
宮城	123,815	1,531	3,039	2,089	2	1	0	4,092	4,860	95	1,128
秋田	57,635	479	1,378	1,042	3	0	0	2,938	2,324	28	1,091
山形	33,075	149	912	864	1	0	0	2,030	1,399	27	252
福島	59,215	604	507	1,663	0	0	0	3,503	3,173	29	553
茨城	70,492	276	1,943	1,746	4	0	0	4,104	3,107	57	920
栃木	33,824	717	1,449	852	2	0	0	1,834	1,762	25	411
群馬	63,221	1,442	3,907	1,567	11	1	0	2,839	3,277	67	926
埼玉	139,901	953	3,185	2,570	8	1	0	4,711	4,481	65	584
千葉	66,406	540	1,246	1,174	7	0	0	2,845	2,456	52	458
東京	115,750	581	1,661	2,065	7	0	0	6,401	4,282	124	503
神奈川	101,895	749	3,534	1,840	6	0	2	6,090	4,902	48	996
新潟	90,476	635	2,406	1,732	11	0	1	3,337	3,213	60	672
富山	74,362	880	2,148	1,698	7	0	0	3,389	3,346	26	488
石川	39,825	443	750	1,199	3	0	0	2,835	2,643	99	614
福井	87,817	436	2,352	1,509	5	0	1	2,942	3,570	34	620
山梨	79,247	344	1,691	1,543	0	0	0	3,873	2,713	41	413
長野	95,842	1,035	4,179	2,557	9	0	0	4,561	4,899	135	796
岐阜	67,610	390	1,464	1,437	11	2	0	3,421	2,598	53	540
静岡	91,526	1,011	2,915	1,778	5	0	0	4,777	3,541	85	669
愛知	88,416	676	2,578	2,012	14	1	1	5,548	3,709	65	698
三重	75,893	1,240	5,883	1,199	5	1	1	3,140	3,647	54	1,007
滋賀	92,969	1,460	3,687	2,610	7	2	0	5,592	6,124	127	1,995
京都	72,824	982	3,393	1,625	17	0	0	3,662	3,917	57	1,074
大阪	95,929	1,205	2,642	1,346	4	2	3	3,413	3,520	80	647
兵庫	53,622	590	1,775	1,306	2	1	0	2,999	2,156	78	829
奈良	71,490	614	2,525	1,546	10	1	0	3,213	3,256	53	978
和歌山	62,970	1,285	2,321	1,104	9	2	1	2,167	2,789	47	658
鳥取	172,689	4,221	10,084	2,043	13	2	1	3,672	10,345	109	4,876
島根	106,708	2,356	2,204	1,948	6	1	0	3,962	4,987	61	1,293
岡山	111,268	1,378	2,818	2,123	12	0	0	3,884	4,743	63	1,182
広島	54,298	528	1,126	1,176	7	1	0	3,631	2,539	34	972
山口	82,116	972	1,405	1,489	15	0	1	4,269	3,315	49	771
徳島	69,015	393	1,133	964	9	1	1	3,145	2,330	66	537
香川	68,657	427	850	1,418	5	1	0	3,803	2,963	29	724
愛媛	58,547	824	2,000	1,156	8	0	0	2,789	2,361	37	616
高知	83,058	539	3,399	1,839	6	0	0	5,704	4,154	75	942
福岡	84,918	631	1,413	1,880	3	0	1	4,521	3,315	73	650
佐賀	97,474	1,320	2,735	1,715	10	1	2	3,975	4,278	71	1,271
長崎	54,180	939	2,044	832	2	0	2	1,438	1,877	22	667
熊本	82,164	1,007	3,159	1,690	7	2	0	3,046	3,746	37	666
大分	94,297	1,236	2,461	1,564	11	0	0	2,822	3,106	44	748
宮城	103,344	2,071	4,695	1,859	7	3	3	4,128	4,786	36	1,115
鹿児島	113,984	708	2,411	2,513	13	0	0	4,988	4,711	71	844
沖縄	155,947	2,664	4,166	2,327	10	1	0	4,718	4,188	80	842
total		45,940	118,583	78,074	322	30	21	180,141	171,457	2,834	43,957

-kara		-node		-tame	
ですから	771	ですので	614	説明のため	681
別紙のとおり決定したから	186	通告がありませんので	594	念のため	566
審査中の事件について左記により閉会中もお継		討論の通告がありますので	593	職務のため	452
続審査を要するものと決定したから	156	議長のお許しをいただきましたので	577	議案調査のため	358
お手元に配布しておきましたから	125	お手元に配付いたしておきましたので	506	県民のため	343
お手元に配付いたしてありますから	117	通告がありますので	402	地方自治法第百二十一条第一項の規定により説明	
左記のとおり決しましたから	110	討論の通告がありませんので	378	のため	287
その後の異動は次のとおりであるから	88	お手元に配付してあります議長報告のとおりであ		何のため	280
成規の賛成を得て少数意見として留保したから	59	りますので	306	県勢発展のため	243
議事日程は当日お知らせいたしますから	59	原案を可決すべきものと決定したので	299	その実現のため	242
配布しておきましたから	55	その処理の経過及び結果の報告を請求すべきもの		復興のため	177
		と決定したので	248		
-okagede		-seide		-bakarini	
皆様のおかげで	7	親方日の丸的な経営感覚と公務員としての身分保		また中に入ったばかりに	2
秋田犬「ゆめ」とシベリア猫「ミール」のおかげで	3	障のせいで	2	充実をさせたばかりに	2
世の中の景気のおかげで	3	橋梁のせいで	2	過度に民間にアクセスされたばかりに	2
その認得のおかげで	3	確かに未曾有の地震と津波のせいで	1	耐震強度が不足した建物が市街地に放置されたば	
知事のおかげで	3	一部の保護者のせいで	1	かりに	1
日本のおかげで	3	よくTPP参加推進論者とと言われる方からの農業		外国人の患者を受け入れたばかりに	1
皆さんののおかげで	3	のせいで	1	懲戒免職になったばかりに	1
ドクターヘリのおかげで	2	ヨない運動のせいで	1	県下一律の制度に固執したばかりに	1
御協力ののおかげで	2	市町村のせいで	1	ブラック企業に入社したばかりに	1
競技会場では語学ボランティアのおかげで	2	異常気象のせいで	1	通年議会も開催されたばかりに	1
		つまりこの制度のせいで	1	やっぱりその取り組みをしなかったばかりに	1
		おまえのせいで	1		

図 3: 根拠のテンプレートにより抽出された節 (Top 10)

-tame		-you	
財産を守るため	235	議員がおっしゃるよう	459
地域の自主性及び自立性を高めるため	211	おっしゃるよう	246
理由 内容及びその重大性よりしてなお審査検討		思うよう	169
を要するため	155	議員おっしゃるよう	85
県民の安全・安心を確保するため	131	許可されるよう	81
命を守るため	126	議会閉会中においても継続して調査並びに審査で	
これを実現するため	110	きるよう	80
これを実現するため	109	今おっしゃるよう	74
こうした課題に対応するため	89	議員のおっしゃるよう	65
県民の命を守るため	86	都民の負託に応えるよう	59
二 理由 いずれもなお調査検討を要するため	84	知事がおっしゃるよう	47
-beku		-to	
議長におかれましてはしかるべく	20	この条例を制定しようと	336
県民の負託に応えるべく	15	所要の改正をしようと	189
しかるべく	15	関係市町村の負担額を変更しようと	147
議長においてしかるべく	12	所要の規定の整理を行おうと	117
県民の期待に応えるべく	8	どのように取り組もうと	112
起こるべく	7	どのように取り組んでいこうと	96
県民の負託にこたえるべく	7	どのように対応しようと	82
県民の皆様の負託に応えるべく	6	どのように取り組まれよう	80
説明責任を果たすべく	6	今後どのように取り組んでいこうと	59
強い経済を取り戻すべく	5	議会の同意を得よう	58

図 4: 目的のテンプレートにより抽出された節 (Top 10)

1 character		2 characters		3 characters		4 characters	
県 (prefecture)	11925	地域 (region)	18551	市町村 (municipalities)	5812	取り組み (approach)	7892
国 (country)	10114	県民 (prefectural people)	15226	活性化 (activation)	4277	子供たち (children)	4042
私 (I)	5481	これ (this)	14972	積極的 (aggressive)	3277	さまざま (various)	3592
今 (now)	4733	対応 (deal with)	14568	これら (these)	3177	それぞれ (each)	3427
人 (person)	4205	確保 (secure)	12802	効果的 (effective)	3067	中小企業 (small and medium enterprises)	1788
何 (what)	2788	必要 (need)	10705	皆さん (everyone)	3054	児童生徒 (schoolchildren)	1773
命 (life)	2567	安心 (peace of mind)	10448	高齢者 (senior citizens)	2577	みずから (oneself)	1421
形 (form)	2350	推進 (promotion)	9856	具体的 (concrete)	2453	持続可能 (sustainable)	1277
力 (power)	2075	実現 (realization)	9486	安定的 (stable)	2041	サービス (service)	1072
場 (place)	1811	今後 (henceforth)	8844	総合的 (comprehensive)	1818	地域住民 (local resident)	1060
5 characters		6 characters		7 characters		8+ characters	
子どもたち (children)	1229	東日本大震災 (Great East Japan Earthquake)	1519	インターネット (Internet)	302	再生可能エネルギー (renewable energy)	846
まちづくり (town planning)	913	行政サービス (administrative services)	582	自然エネルギー (natural energy)	220	コミュニケーション (communication)	250
地方自治法 (Local Autonomy Act)	844	外国人観光客 (foreign tourists)	415	パラリンピック (Paralympic)	218	東京オリンピック (Tokyo Olympic)	240
エネルギー (energy)	790	プロジェクト (project)	334	エネルギー政策 (energy policy)	195	地域包括ケアシステム (community-based integrated care systems)	217
教育委員会 (Board of Education)	716	介護サービス (care service)	333	リーダーシップ (leadership)	155	南海トラフ巨大地震 (Nankai megathrust earthquake)	142
地域づくり (community development)	605	地方公共団体 (local public body)	328	インセンティブ (incentive)	154	市町村教育委員会 (Municipal Board of Education)	130
地域活性化 (Regional activation)	487	ネットワーク (network)	288	地球温暖化対策 (global warming countermeasure)	154	コミュニケーション能力 (communication ability)	122
農林水産業 (Agriculture, forestry and fisheries)	457	オリンピック (Olympic)	269	イメージアップ (image enhancement)	154	リニア中央新幹線 (Linear Chuo Shinkansen)	117
百二十一条 (article 121)	443	社会保障制度 (social security system)	267	イノベーション (innovation)	148	東日本大震災津波 (Great East Japan Earthquake Tsunami)	115
健康づくり (Health promotion)	377	県内中小企業 (small and medium enterprises in the prefecture)	267	モチベーション (motivation)	131	ボランティア活動 (volunteer activity)	99

図 5: 根拠と目的のテンプレートにより抽出されたキーワード

- [5] 伊木惇, 亀井清華, 藤田聡. レビューを対象とした信頼性判断支援システムの提案. 情報処理学会論文誌, vol. 55, no. 11, pp. 2461–2475, (2014).
- [6] 佐藤美沙, 柳井孝介, 柳瀬利彦, 是枝祐太, 丹羽芳樹. 国会会議録を用いたディベート人工知能による意見生成. 第 31 回人工知能学会全国大会, 4Q1-9in2, (2017)
- [7] 松本拓也, 渋谷英潔, 森辰則. 情報信憑性判断支援のための対話型条件結論マップ生成に向けた条件と結論の抽出. 言語処理学会第 20 回年次大会発表論文集, pp. 262–265, (2014).
- [8] 大塚淳史, 関洋平, 神門典子, 佐藤哲司. コミュニティQA を用いたクエリ拡張のためのコンテキスト抽出に関する一考察. 日本データベース学会論文誌, Vol. 11, No. 1, pp. 1–6, (2012).
- [9] 小柴等, 森川想. 議事録を用いた我が国における議会・行政の関係性分析手法. 人工知能学会論文誌, vol. 34, no. 5, pp. E-J47-1-10, (2019).
- [10] 大南勝, 掛谷英紀. 国会会議録に基づく短命大臣の特徴分析 第 2 報. 言語処理学会第 24 回年次大会発表論文集, pp. 1167–1170, (2018).
- [11] 関喜史. 世代による政治ニュース記事の閲覧傾向の違いの分析. 第 32 回人工知能学会全国大会, 3O2-OS-1b-04, (2018)
- [12] 中川侑, 武田拓也, 吉元涼介, 芳鐘冬樹. ランダムフォレストを用いた国会会議録のイデオロギー分析. 情報科
- 学技術フォーラム (FIT) 講演論文集, vol.15, no.4, pp. 277–278, (2016).
- [13] 庵功雄, 高梨信乃, 中西久美子, 山田敏弘. 中上級を教える人のための日本語文法ハンドブック. 白川博之 (監), 株式会社スリーエーネットワーク, (2001).

SIGIR2019 参加報告

Report on SIGIR2019

野本 昌子^{1*} Bhattacharjee Anupam¹

¹ ヤフー株式会社

¹ Yahoo Japan Corporation

Abstract: The 42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019) took place in Paris, France in July 2019. A total of 426 papers for the full paper track were submitted and reviewed with an acceptance rate of 19.7%. About 73.8% of the accepted papers were related to neural network, distributed over about 95.5% of technical sessions.

1 はじめに

SIGIR¹は情報検索 (Information Retrieval, 以下 IR) 分野のトップカンファレンスで, ACM SIGIR²が後援している国際会議の中でも歴史が古く, 1978年から毎年開催されている。従来は欧米で開催されていたが, 近年はアジア・オセアニア, 南北アメリカ, ヨーロッパを順に回っている [1]。周辺分野には自然言語処理, 人工知能, データマイニング, ヒューマン・コンピュータ・インタラクション等がある [2]。

本稿では SIGIR2019 の参加報告として, 概要, 動向, 論文の一部と国内のコミュニティの動向を紹介する。

2 SIGIR2019

2.1 概要

SIGIR2019 はパリのシテ科学産業博物館で, 7月21日から25日まで開催された。表1³に SIGIR2016-2019 の開催概要を示す。今回のプログラムも例年通りチュートリアルとワークショップが各1日, 本会議が3日間であった。

参加者数は SIGIR2012 から SIGIR2015 にかけては減少傾向にあったが ([3] p.9), SIGIR2016 から盛り上がりを見せ, 日本開催の SIGIR2017 で 900 名を超えたが, 今回は 1,025 名とさらに記録を更新した。

*連絡先: ヤフー株式会社
千代田区紀尾井町 1-3
E-mail: mnomoto@yahoo-corp.jp

¹SIGIR Conference on Research and Development in Information Retrieval

²検索・情報アクセス分野の国際学会 ACM の IR の分科会 SIGIR(Special Interest Group on Information Retrieval)。本稿で国際会議 SIGIR との区別のため, 以降, 'ACM SIGIR' と記す。

³表中, フルペーパー, ショートペーパー, デモ, SIRIP の数字は採択数または採択数/投稿数 (採択率) を示す。

前回のスポンサーは4カ国の企業で, 米国のIT系, 小売/eコマースと中国のIT系の企業が多かった [4]。今回のスポンサー (表2⁴) は9カ国に増えたが, 中国の企業は大幅に減少した。米国企業はあまり変化がないが, アップルがプラチナスポンサーとして加わった。

運営に関する新しい試みとしてはイベントアプリ Whova⁵の使用が参加者に推奨され, 会期中のスケジュール管理やコミュニケーション等のツールとして約1,200アカウント⁶で利用された。

SIGIR2019 のワードクラウド (図1⁷) では Search(検索) より Recommendation(推薦) や Learning(学習) 等が目立つ。

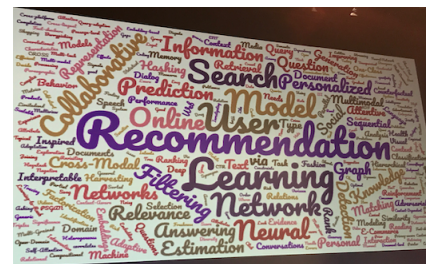


図1: SIGIR2019 ワードクラウド

2.2 投稿・採択状況

図3に SIGIR のフルペーパーの投稿数, 採択数, 採択率の推移 (1999~2019年) を示す⁸。今年は昨年を上

⁴[5] pp.xlii-xlv, [6] pp.xxxviii-xxxix, [7] pp.xxxv-xxxvi, [8] pp.xvii-xviii を元に作成した。SIGIR2016 から SIGIR2018 のスポンサーに該当する場合, 表中に'*' を付している。

⁵<https://whova.com/>

⁶ただし, 企業等の組織アカウントも含む。

⁷オープニングの Program Chair's Welcome のスライドより。

⁸[5]Publication, Overall Acceptance Rate の表より作成した。

表 1: SIGIR2016-2019 の概要

会議(回)	SIGIR2019(42)	SIGIR2018(41)	SIGIR2017(40)	SIGIR2016(39)
会期	July 21-25	July 8-12	Aug 7-11	July 17-21
開催地	Paris, France	Ann Arbor Michigan, USA	Tokyo, Japan	Pisa, Italy
参加者数	1,025	740	911	566
フルペーパー	84/426 (19.7%)	86/409 (21.0%)	78/362 (21.5%)	62/341 (18.1%)
ショートペーパー	108/443(24.4%)	98/327 (30.0%)	121/398 (30.4%)	104/339 (30.7%)
デモ	21/46(45.7%)	18/36 (50.0%)	17/36 (47.2%)	21/35 (60.0%)
チュートリアル	11	11	8	12
ワークショップ	8	10	8	7
SIRIP(Industry Track)	9/12 (75.0%)	9/16 (56.3%)	5	12

表 2: SIGIR2019 のスポンサー

国	分類	企業	2018	2017	2016
アメリカ	IT	Microsoft	*	*	*
		IBM	*	*	*
		Google	*	*	*
		Facebook	*	*	*
		Bloomberg	*		*
		Apple			*
		ebay	*	*	*
小売/e コマース		amazon	*	*	*
韓国	IT	NAVER	*	*	
中国	IT	HUAWAI		*	*
日本	IT, 小売/e コマース	Rakuten		*	
インド	IT	ShareChat			
オーストラリア	IT	Appen			*
スウェーデン	音楽	Spotify	*		
フランス	IT	Criteo AI Lab			
		Technicolor			
ドイツ	e コマース	Zalando			

表 3: フルペーパーの国別分布の推移

会議	1 位	2 位
SIGIR2019	中国 (43%)	米国 (26%)
SIGIR2018	中国 (34%)	米国 (30%)
SIGIR2017	米国	中国
SIGIR2016	米国 (34%)	中国 (23%)

回る 426 本のフルペーパーが投稿されたが採択数は昨年並みの 84 本で、採択率は 19.7%とやや難化し(昨年は 21.0%), SIGIR1999 から SIGIR2019 の平均採択率 19.3%と同程度である。

国別での投稿・採択状況は SIGIR2017 から上位で変化が見られる。中国は SIGIR2017 で投稿数はトップであったが、採択数で米国に及ばなかった([3] p.11)。しかし、SIGIR2018 から採択数でも首位に立ち、今回はさらに米国との差を広げている(表 3⁹)。また、SIGIR2019 の国別投稿数、採択数の分布(図 2¹⁰)を元に採択率を比較すると、中国は 25.5%(採択数 36 件、投稿数 141 件)で米国の 22.2%(採択数 22 件、投稿数 99 件)を上回っている。さらに、フルペーパーの著者の国別分布(SIGIR 2016-2019)(図 4¹¹)でも、2016 年の時点では米国がトッ

⁹[5]p.iii, [6]p.iii, [9]p.6, [8]p.v より作成した。

¹⁰オープニングの Program Chair's Welcome のスライドより。

¹¹[6]p.iii,[8]p.v を元に作成した。

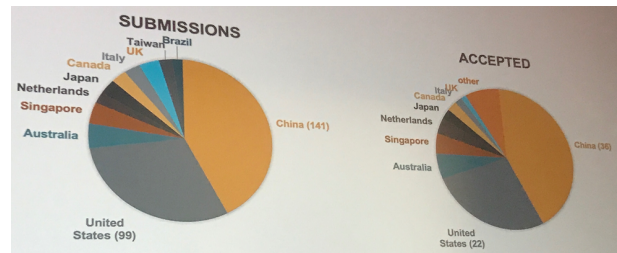


図 2: SIGIR2019 国別投稿数、採択数の分布

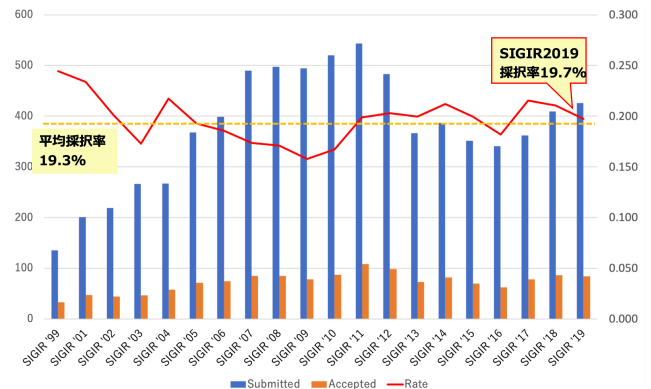


図 3: フルペーパーの投稿数、採択数、採択率の推移 (1999-2019 年)

プ(25%)で、中国は 2 位(23%)であったが、2017 年に逆転し、回を重ねるごとに割合を増やしている。

次に日本の投稿・採択状況については、今回は投稿数、採択数とも 6 位である(図 2)。日本人の著者を含むフルペーパーの採択例(SIGIR2016~2019)を表 4¹²に示す。SIGIR2019 では少なくとも 3 本が採択されている。日本の組織についてはアカデミアに偏っており、早稲田大学、筑波大学、京都大学から採択されている。

SIGIR 全体でのフルペーパーの著者の所属グループ

¹²SIGIR2019 については著者名から日本人らしいと主観で判断した人名を選んだ。他の回について網羅的な調査はしていない。また、TOIS (ACM Transactions on Information Systems) のジャーナル論文の著者は含めていない。

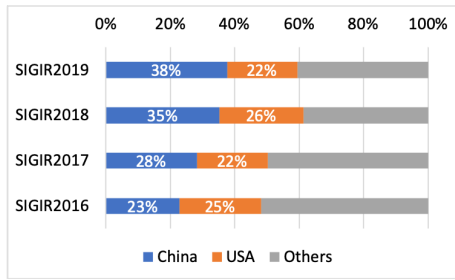


図 4: フルペーパーの著者の国別分布 (SIGIR'16-'19)

表 4: 日本人著者のフルペーパー採択例 (SIGIR'16-'19)

年 (Session)	組織:著者 (*:筆頭)
2019 (7A)	Tetsuya Sakai*(早稲田大)
	(2C) Makoto P. Kato(筑波大), Masatoshi Yoshikawa(京都大)
	(1A) Kenta Takatsu(Cornell 大)
2017 (1A)	Tetsuya Sakai*(早稲田大)
	(4A) Tetsuya Sakai(早稲田大)
	(6B) Sosuke Shiga*, Hideo Joho(筑波大)
2016 (1A)	Tetsuya Sakai*(早稲田大)
	(5B) Tetsuya Sakai(早稲田大)

の分布 (SIGIR2016~2019) (図 5¹³) はアカデミアが多いが, SIGIR2017 以降, 企業の著者も少しずつ増えている。

フルペーパーの企業別著者数 (SIGIR2016~2019) (図 6¹⁴) を見ると, 上位を除き, 全体として年による入れ替わりが激しく, 継続的に採択されている企業は少ない。今回の上位は Alibaba, Microsoft, Google の順で, Microsoft は SIGIR2016 以降, 常に 1 位か 2 位で安定しており, Alibaba と Google は今回急増している。

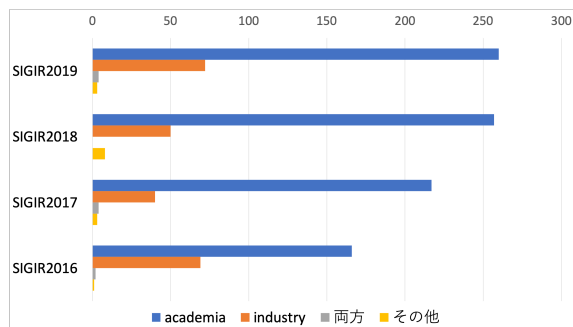


図 5: フルペーパーの所属グループ別著者数 (SIGIR2016-2019)

¹³[10] の full paper の著者の所属, [11], [12], [13] のプログラムの著者の所属を元にアカデミア, 企業, 両方, それ以外に分類した。
¹⁴[10] の full paper の著者の所属, [11], [12], [13] のプログラムの著者の所属を元に作成した。

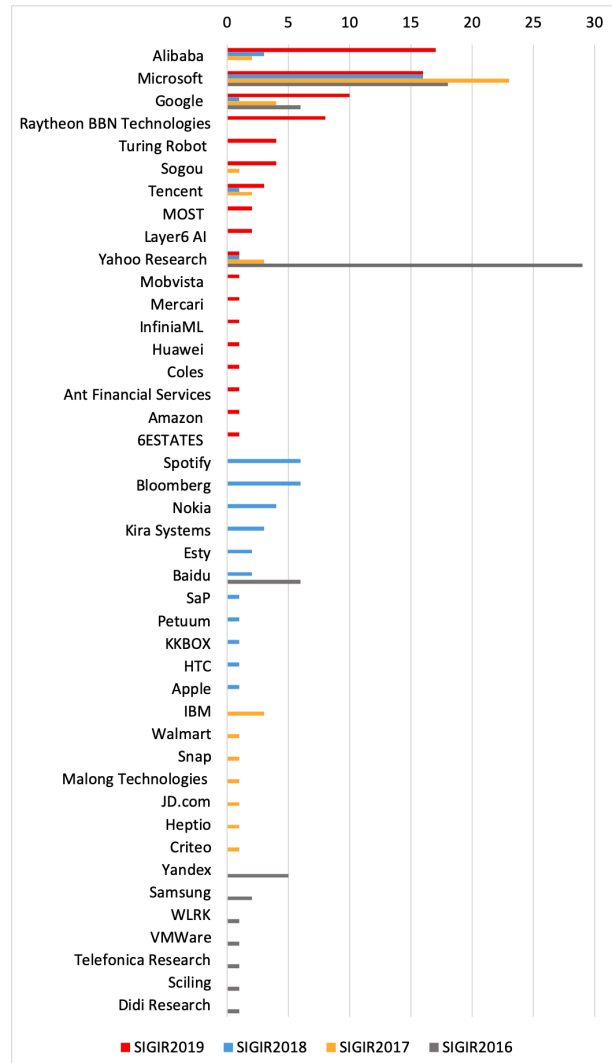


図 6: フルペーパーの企業別著者数 (SIGIR2016-2019)

2.3 トラック

SIGIR2019 のトラック構成を表 5 に示す。今回は (g) Evaluation のトラックが追加された¹⁵。今回はトラック別に投稿, 選考が行われたが, 今回は 1~2 トラックを選択して投稿が行われ, 選考は投稿論文全体で行われた¹⁶。

SIGIR2019 のトラック別投稿・採択状況を図 7 に示す。今回は (d) コンテンツ分析・推薦・分類, (a) 検索とランキングに関連する論文の投稿が多く, (d) は前回も最も投稿数が多い (124 本, 全投稿の 30%) トラックであった ([6] p.iii)。今回, 関連する論文の採択が多かったトラックも (d) と (a) だが, 前回採択数が多かったト

¹⁵他にトラック名の変更があり, (b) は 2018 年は Search and Ranking (Core IR), (d) は Content Recommendation, Analysis and Classification であった。

¹⁶したがって SIGIR2018, SIGIR2019 のトラック別論文の分布は単純に比較できない。

表 5: SIGIR2019 のトラック構成

- (a) Search and Ranking
- (b) Future Directions
- (c) Domain-Specific Applications
- (d) Content Analysis, Recommendation and Classification
- (e) Artificial Intelligence, Semantics, and Dialog
- (f) Human Factors and Interfaces
- (g) Evaluation

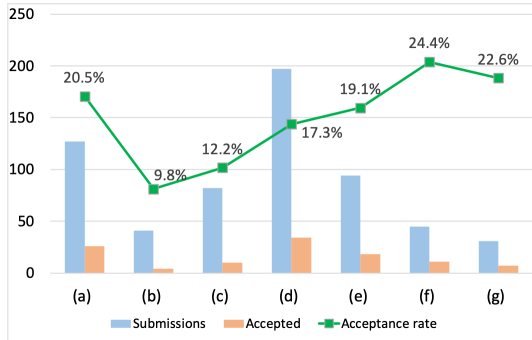


図 7: トラック別投稿・採択状況 (SIGIR2019)

トラックは (c) 特定分野の応用 (20 本, 全採択の 23%), (f) Human Factors とインターフェイス (20 本, 同 23%) であった ([6] p.iii).

2.4 セッション

2019 年のセッション別論文数¹⁷は推薦 (3A,7C), 適合性と評価 (7A,9B) が順に多い。推薦に関する論文が多い傾向は 2016 年から続いており [4], 今回の推薦のセッションの論文を見ると RecSys, WWW 等, 周辺分野も含め他の国際会議の参照が多く, 研究コミュニティ間の融合が進んでいるように見受けられる。また, 新しいセッションとしては解釈性と説明性に関するもの (3B), 昨年, New IR Application のセッションで研究例が見られたファッションマッチ (9A), クロスモーダル検索に関するもの (7B) 等がある。

前述のイベントアプリ Whova 上の各セッションの参加登録数と like 数の分布を図 8 に示す¹⁸。参加登録数の多かったセッションは順に 7A (適合性と評価 1), 1C (検索意図), 3B (解釈性と説明性), 1A (ランキング学習 1), 4B (クエリ), 2C (知識とエンティティ) である。like 数が特に多かったのは 2C, 続いて 1C, 7A で, like 数が 20 件以上のセッションは 7C (推薦 2) を除き, 参加登録数が 100 件以上であった¹⁹。

¹⁷ 同じ名前のセッションがプログラム上複数の時間に分かれている場合はまとめて 1 セッションとカウントした。

¹⁸ アプリの使用は推奨に留まり必須ではなかったため, 参加登録数, like 数は参考値である。

¹⁹ アプリ上で like は参加登録の有無とは独立にマークできる。like 数と参加登録数の積率相関係数は 0.79。

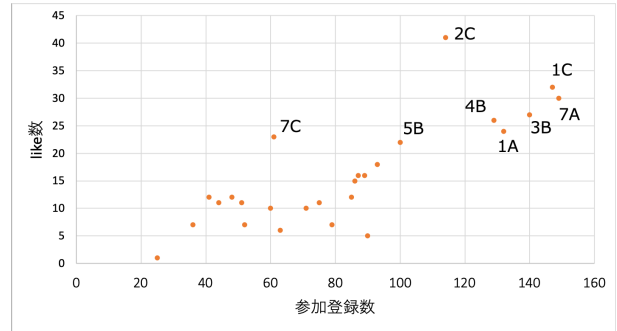


図 8: アプリ (Whova) 上のセッション別参加登録数と like 数の分布 (SIGIR2019)

2.5 ニューラルネットに関する動向

SIGIR2016 の基調講演で Christopher Manning 氏 (スタンフォード大学) は SIGIR にも深層学習が急速に浸透すると予想し ([14], p.72), SIGIR2017 の Neu-IR²⁰ のオープニングでは約 4 分の 1 の論文が neural IR に関するものであり, 「ニューラルネットの年 (The year of the Neural Nets)」と報告された。

そこで, SIGIR2019, SIGIR2018 のフルペーパー²¹を対象にニューラルネット (以下, NN と略す) に関する論文かどうか調べた²²。図 9, 図 10 に各々セッション単位のフルペーパーの論文数と NN 関連の論文数を示す。

SIGIR2018 のフルペーパーは約 46.0% (40/87 本) が NN 関連であったが, SIGIR2019 では約 73.8% (62/84 本) と大幅に増加した。

セッション単位では, SIGIR2018 では約 73.1% (19/26) のセッションに NN 関連の論文が分布していたが²³, SIGIR2019 では約 95.5% (21/22) に広がっている。SIGIR2019 のセッションで NN 関連の割合が大きいものは推薦 (3A, 7C), 質問応答 (2A), 対話 (5A), ファクトチェック, プライバシーと Legal (3C), Personalization と個人情報の検索 (6B), 多言語とクロスモーダル検索 (7B), ファッションマッチ (9A), Hashing (8B) 等である。

SIGIR2018 で NN 関連の割合が大きかったセッションは New IR Applications, Conversational Systems, Question Answering, Hashing & Embedding であった。このうち New IR Applications のセッションの論文はク

²⁰ Neural Information Retrieval のワークショップ。https://neu-ir.weebly.com

²¹ テクニカルセッションのフルペーパーのうち, TOIS (ACM Transactions on Information Systems) のジャーナル論文を除く。

²² 該当するかどうかの判断は基本的にはキーワード文字列 ('neural net', 'deep learning' 等) ベースで行った。future work のみに出現するなど, 出現位置から非該当と容易に判別できるものは除いたが, NN の手法と提案手法の比較等は NN 関連としている。

²³ 同じ名前のセッションがプログラム上複数の時間に分かれている場合はまとめて 1 セッションとカウントした。

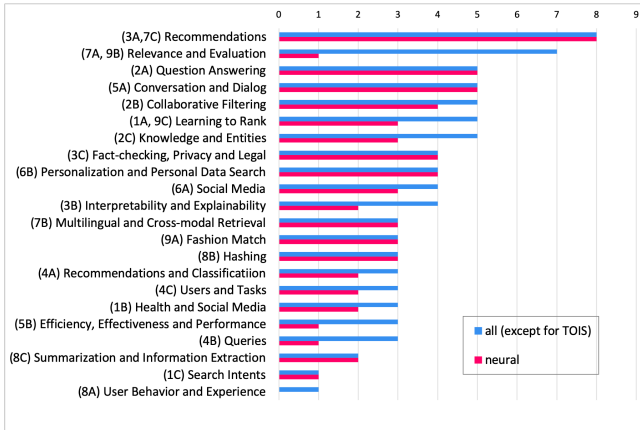


図 9: セッション別のフルペーパーとニューラルネット関連の論文数 (SIGIR2019)

ロスモーダル検索と分野特有の知識を利用したマッチングに関するものであり、これらに対応する SIGIR2019 のクロスモーダル検索関連 (7B) とファッションマッチ (9A) のセッションの論文も全て NN 関連である。また、その他の対話関連、質問応答、Hashing 関連のセッションについても同様に今回も NN 関連が多い²⁴。

特に NN 関連の論文の増加が目立つのは推薦 (3A, 7C) のセッションである。また、Knowledge Base や Entities 関連 (2C) でも NN 関連の研究が増えている。

NN 関連の論文の手法に関するキーワードでは attention が頻出するが (約 40 件²⁵)、attention mechanism と明記されているものは約 18 件、BERT は約 4 件、Transformer が少数あった。他の手法では RNN 系が多く、RNN, GRU, LSTM が各 15 件程度²⁶、Seq2Seq が約 3 件あった。次に多かったのは CNN 系で、CNN が約 14 件、VGG と GCN が各 2,3 件あり、他には GAN が約 2 件あった。

2.6 受賞論文

Best paper は Wang ら [16] のオンラインのランキング学習に関する研究で、Dueling Bandit Gradient Descent (DBGD) を改良し、クエリの重みの異なるランキングから合成した検索結果をユーザに提示し、ユーザからのクリックを元にモデルを更新する際、ユーザの実際の評価対象から外れたものに起因するノイズを除去して勾配予測の variance を小さくする手法を提案した。

²⁴SIGIR2019 のセッションでは各々 5A, 2A, 8B に対応する。

²⁵調査は論文中のキーワード文字列の出現を調べたもので、手法の正確な分布ではない。

²⁶各々 Bi-directional RNN, Bi-GRU 各数件, Bi-LSTM 約 8 件を含む。

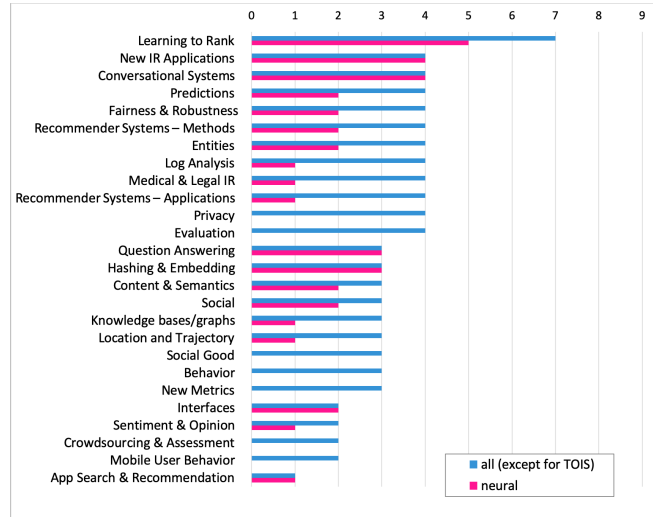


図 10: セッション別のフルペーパーとニューラルネット関連の論文数 (SIGIR2018)

Test of Time Award²⁷は、情報検索の評価指標 n-DCG の改良版 (α -nDCG) を提案した Charles Clarke ら [17] に贈られた。

2.7 基調講演

マサチューセッツ工科大学のクロフト教授は情報検索におけるインタラクションの重要性について講演し [18]、ユーザーインタラクションに関する情報検索研究の歴史を振り返り、質問応答や対話的検索におけるインタラクションや研究例を紹介した。

フランス国立情報学自動制御研究所の Research Director のシュミド氏は、目に見える世界の自動理解というテーマで講演した [19]。合成データを用いた人体のパーツの認識や、動画と音声を用いた VideoBERT で対応づけたマルチモーダルデータを用いた弱教師学習で、料理ビデオ中のアクションを分類した例、時間依存性 (temporal dependency) や人とオブジェクトの関係のグラフを用いたビデオ中のアクションの検出例、Graph Neural Networks や RPN (Region Proposal Network)、attention 等を用いた行動予測の研究等を紹介した。今後の研究の方向性としては現実世界とのインタラクションからの学習を挙げた。

2.8 チュートリアル

今回は表 6 に示す 11 のチュートリアルがあった²⁸。

²⁷過去 10-12 年の ACM SIGIR がスポンサーの会議で最もインパクトの大きかった論文に贈られる賞。

²⁸<http://sigir.org/sigir2019/program/tutorials/>

表 6: Tutorial (SIGIR2019)

Tutorial	トピック
(a) Building Economic Models and Measurement of Search	検索の経済モデルと評価指標, 情報フォレンジング理論等.
(b) Effective Online Evaluation for Web Search	Web 検索のオンライン評価.
(c) Learning to Rank in theory and practice	ランキング学習の理論と実際, 勾配ブースティング, NN 応用, Unbiased Learning 等.
(d) Deep Chit-Chat	Deep Learning のチャットボットへの応用.
(e) Example-driven Search	example ベースの検索.
(f) Explainable Recommendation and Search	説明可能な推薦と検索.
(g) Web Table Extraction, Retrieval and Augmentation	Web テーブル抽出, 検索, 拡張.
(h) Deep Natural Language Processing for Search Systems	産業界からみた検索システム向けの Deep Learning ベースの自然言語処理.
(g) Extracting, Mining and Predicting Users' Interests from Social Networks	ユーザの意図のマイニング, ソーシャルネットワークからのマイニング, 評価等.
(h) Fairness and Discrimination in Retrieval and Recommendation	検索, 推薦における Fairness(公平性).
(i) Learning to Quantify	Quantification(定量化).

Yandex と Facebook による (b) Web 検索のオンライン評価のチュートリアル [15] では, オンライン評価指標の構築等の他に, Yandex でのオンライン実験の実例を取り上げ, オンラインテスト実施の重要性, 限界やバイアス等の問題など, 得られた教訓を紹介した.

2.9 SIRIP

SIRIP²⁹は, 企業での IR の応用に関するもので, 今回は 4 セッションが行われた³⁰.

Netflix からはストリーミングサービスのコンテンツ検索の事例紹介があった [20]. Netflix 上の検索には 'Fetch' (検索するタイトルが決まっている場合), 'Find' (特定の俳優等, 要求は明確だが該当するタイトルが不明の場合), 'Explore' (ジャンル名等, より広範な条件で探したい場合) の 3 種類がある. 検索の少なくとも 13% は利用できないコンテンツに関するものと推定され, その場合にどのような推薦が望ましいか, また検索結果と推薦をどう組み合わせるか等, 課題が多い. 入力時のユーザの負担を軽減するにはインスタントサーチが有効であるが, 検索クエリは Web 検索に比べて文字数も語数も少なく, 検索意図の解釈が難しい. また, クエリ入力時のキーストロークを少なくしたいが, ハングルの入力等, システムによってはクエリより文字数が増える場合もある. ユーザの行動データを重視し, 従来の情報検索の範囲を超えたアプローチが必要とのことである.

2.10 ワークショップ

今回は表 7 に示す 8 ワークショップがあった³¹. 新しいワークショップは, 対話エージェントに関するもの (WCIS2019), 情報の事実性に関するもの (ROME2019), IR における Fairness, Accountability, Confidentiality, Transparency, Safety に関するもの (FACTS-IR

2019) で, 全体として事実性や倫理問題に関するトピックを扱うものが多くなっている.

対話エージェントとインタラクションに関するワークショップ WCIS2019³²では Facebook から AI ボット同士の交渉の研究が紹介された. 複数の種類の品物をどう分けるか交渉するが, 相手のボットにとっての商品の価値は知らされず, 対話から推測する. 対話の応答候補をいくつか生成しておき, その後の展開を推定して最終的にもっとも reward(報酬)が高くなる発話を選ぶ, dialogue rollouts という手法が有効である.

2.11 SIGIR2019 のトレンド

今回のトレンドを以下にまとめる.

- NN 関連の論文がさらに増加 (フルペーパーの約 73.8%, テクニカルセッションの約 95.5%).
- 国別では中国が躍進中. 投稿数, 採択数とも首位.
- トラック別投稿数は前回に続きコンテンツ分析・推薦・分類が多く, 続いて検索とランキングが多い.
- セッション別論文数は推薦, 適合性と評価が多い.
- 新しいセッションは解釈性と説明性 (3B), ファッションマッチ (9A), クロスモーダル検索関連 (7B) 等.
- 参加登録の多かったセッションは適合性と評価 (7A), 検索意図 (1C), 解釈性と説明性 (3B), ランキング学習 1(1A), クエリ (4B), 知識とエンティティ (2C) 等.
- 新しいワークショップは対話エージェント関連, 情報の事実性や IR の倫理問題に関するもので, 全体として後者に関するトピックが増えている.

2.12 論文紹介

前述のアプリで参加登録数または like 数の多かったセッションから一部の論文を紹介する.

²⁹SIGIR Symposium on IR in Practice

³⁰SIGIR2016, SIGIR2017 では各 3 セッション, SIGIR2018 では 6 セッションであった.

³¹<http://sigir.org/sigir2018/program/workshops/>

³²<https://sites.google.com/view/wcis/>

表 7: Workshop (SIGIR2018)

Workshop	回	トピック
WCIS2019	1	対話エージェント, インタラクションシステム.
FACTS-IR 2019	1	IR における Fairness(公平性), Accountability, Confidentiality, Transparency(透明性), Safety(安全性).
EARS 2019	2	説明可能な推薦と検索.
BIRNDL 2019	4	ビプリオメトリクスによるデジタルライブラリの検索と自然言語処理.
OSIRRC 2019	3	ad hoc 検索のテストコレクションを用いた replicability チャレンジ.
NewsIR '19	3	ニュース情報検索. データジャーナリズム, フェイクニュース, ファクトチェック, ニュース推薦等.
ROME 2019	1	ファクトチェック, 事実と異なるニュースの検出, ソーシャルメディアへの誤情報拡散の分析等.
ECOM '19	3	e コマース.

2.12.1 7A: Relevance and Evaluation 1

Sakai ら [21] は評価指標自体の評価として, adhoc 検索の指標と diversified IR(多様化検索)の指標(計 30)を対象にユーザの preference との相関を調べた. 実験に参加した 15 名に対してクエリと 2 種類の検索結果を個別に提示し, 適合性については全体としてクエリにより適合している検索結果, 多様性についてはより多くのユーザを満足させる検索結果を選択させる実験を行った. 参加者の preference と指標の一致を調べたところ, 適合性については adhoc 検索でよく用いられる指標(例: nDCG)はよく合っていたが, 多様性の指標では意図意識型(intent-aware)の指標の結果が良くなかった.

2.12.2 1C: Search Intents

Zhang ら [22] は, ユーザの検索意図を捉える分散表現空間を学習する GEN Encoder を提案した. クエリは検索意図を表すには不十分な場合もあるが, クリックが似ているクエリは意図も近いと考え, クリックを検索意図を表す implicit なフィードバックとして学習し, クリックが同じクエリが近くなるように分散表現空間にマッピングする. さらにマルチタスク学習で精度を高めている.

2.12.3 2C: Knowledge and Entities

Imrattanatrai ら [23] はテキストからのエンティティのプロパティ集合の認定を knowledge graph(以下 KG と略す)を利用してゼロショット学習で行う方法を提案した. 例えば歌手のエンティティについてその作品(アルバム, トラック)との関係を表すプロパティ('album', 'track')を認定する場合, KG の各プロパティに相当する情報を含む連結トリプルから, 連結述語に対する主語(歌手)と目的語(アルバム, トラック)を抽出し, これらを含む文を学習に利用する. KG のプロパティの埋め込み方法は KG の異なる部分を用いて比較した. さらにエンティティのタイプを用いて精度を高めている.

2.12.4 3B: Interpretability and Explainability

Balog ら [24] は transparency(透明性)と scrutability を備えた推薦システムを提案した. ここで scrutability とはシステムが認識したユーザの preference をユーザ自身が直接編集して修正できることである. 提案システムは transparency を保ちながら従来システムと同等の性能である. また, scrutability についてはシステムの出力した preference とユーザの認識が異なる場合が多く, ユーザによる修正で改善の余地があることを示した.

3 国内のコミュニティの動向

2017 年 12 月に ACM SIGIR の地方支部である東京支部³³が誕生し, 国内外の研究コミュニティの交流の促進のため, SIGIR 関連のセミナー等のイベントや会員向けの情報発信を行っている. 2019 年 10 月には今回の SIGIR2019 を含む情報アクセス系の国際会議の論文の読み会(IR Reading2019 秋)³⁴を開催した. また, 情報検索研究における日本のプレゼンス向上のため, 同分野の国際会議に投稿予定の論文を対象にしたメンタリングワークショップ POWIR を開催しており, 今回は論文 2 本が SIGIR2019 で採択された. 次回の POWIR 2020 Winter³⁵は 2020 年 1 月に開催予定である.

4 おわりに

IR のトップカンファレンス SIGIR2019 が 2019 年 7 月にパリで開催され, 過去最多の 1,025 名が参加した. フルペーパーは昨年を上回る 426 本が投稿され, 採択率は 19.7%とやや難化した(昨年は 21.0%). 国別では投稿数, 採択数とも中国の首位が続く. NN に関連する研究はテクニカルセッションの約 95.5%, フルペーパーの約 73.8%に広がった. 全体として推薦に関する論文が多い傾向が続いている.

³³Tokyo ACM SIGIR Chapter, <http://sigir.jp/>

³⁴http://sigir.jp/post/irreading_2019fall/

³⁵<http://sigir.jp/post/2019-10-17-powir2020winter/>

SIGIR2020³⁶は2020年7月25日から30日まで中国の西安で開催される。ACM SIGIR 東京支部ではSIGIR2020等の投稿論文のメンタリングワークショップPOWIRを2020年1月に開催する。SIGIR2020では日本からさらに多くの投稿と採択があることを願っている。

参考文献

- [1] SIGIR: List of SIGIR Locations and History, <http://sigir.org/general-information/history/> (accessed Nov 9, 2019)
- [2] Toshihiro Kamishima: ML, DM, and AI Conference Map, <http://www.kamishima.net/archive/MLDMAImap.pdf> (accessed Nov 9, 2019)
- [3] 櫻 惇志: SIGIR2017 参加 (開催) 報告-会議概要とニューラルネットワークな情報検索-, 第 11 回テキストアナリティクス・シンポジウム, <http://www.ieice.org/~nlc/attachment/NLC20170908-SIGIR-2017-Keyaki.pdf> (2017)
- [4] Masako Nomoto, Riku Togashi, Shogo Suzuki, Hiroaki Shiino: SIGIR2018 参加報告, 第 20 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会 (SIG-AM-20-06), pp.36-43 (2018)
- [5] Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, <https://dl.acm.org/citation.cfm?id=3331184> (2019)
- [6] The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, <https://dl.acm.org/citation.cfm?id=3209978> (2018)
- [7] Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, <https://dl.acm.org/citation.cfm?id=3077136> (2017)
- [8] Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, <https://dl.acm.org/citation.cfm?id=2911451> (2016)
- [9] Arjen de Vries, Hang Li, Ryen White: ACM SIGIR 2017 - Opening - PC Chairs, <https://www.slideshare.net/arjenpdevries/acm-sigir-2017-opening-pc-chairs> (2017)
- [10] ACM SIGIR 2018 Conference Organization: ACM SIGIR 2018 Accepted Papers, <http://sigir.org/sigir2018/accepted-papers/> (2018)
- [11] SIGIR2019, <http://sigir.org/sigir2019/> (2019)
- [12] SIGIR2017, <http://sigir.org/sigir2017/> (2017)
- [13] SIGIR2016, <http://sigir.org/sigir2016/> (2016)
- [14] Christopher Manning: Natural Language Inference, Reading Comprehension and Deep Learning, <https://nlp.stanford.edu/manning/talks/SIGIR2016-Deep-Learning-NLI.pdf> (2016)
- [15] Alexey Drutsa, Greb Gusev, Eugene Kharitonov, Denis Kulemyakin, Pavel Serdyukov, Igor Yashkov: Effective Online Evaluation for Web Search, <https://research.yandex.com/tutorials/online-evaluation/sigir-2019> (2019)
- [16] Huazheng Wang, Sonwoo Kim, Eric McCord-Snook, Qingyun Wu, Hongning Wang: Variance Reduction in Gradient Exploration for Online Learning to Rank, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.835-844 (2019)
- [17] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon: Novelty and diversity in information retrieval evaluation, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp.659-666 (2008)
- [18] W. Bruce Croft: The Importance of Interaction for Information Retrieval, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, p.1 (2019)
- [19] Cordelia Schmid: Automatic Understanding of the Visual World, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, p.3 (2019)
- [20] Sudarshan Lamkhede, Sudeep Das: Challenges in Search on Streaming Services: Netflix Case Study, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.1371-1374 (2019)
- [21] Tetsuya Sakai, Zhaohao Zeng: Which Diversity Evaluation Measures Are "Good"?, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.595-604 (2019)
- [22] Hongfei Zhang, Xia Song, Chenyan Xiong, Corby Rosset, Paul N. Bennett, Nick Craswell, Saurabh Tiwary: Generic Intent Representation in Web Search, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.65-74 (2019)
- [23] Wiradee Imrattana-trai, Makoto P. Kato, and Masatoshi Yoshikawa: Identifying Entity Properties from Text with Zero-shot Learning, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.195-204 (2019)
- [24] Krisztian Balog, Filip Radlinski, Shushan Arakelyan: Transparent, Scrutable and Explainable User Models for Personalized Recommendation, Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.265-274 (2019)

³⁶<http://sigir.org/sigir2020/>