情報処理学会論文誌　データベース　Vol.11　No.3　12–22 (Oct. 2018)

職業ごとの行動に関する知識の収集

馬縣 美穂[1,a)]　笹野 遼平[2]　高村 大也[1,3]　奥村 学[1]

受付日 2018年3月8日, 採録日 2018年7月6日

概要：本研究では，ある職業の人間がとる行動を獲得するためのシステムを提案する．提案システムは，対象の職業が主語となっている文から行動を抽出する主語ベース部，および，対象の職業に従事するユーザによって書かれた文から本人の行動を抽出する著者ベース部の 2 つの要素で行動を収集し，得られた行動と職業の間のカイ二乗値を計算することで職業に特徴的な行動を獲得する．クラウドソーシングを用いた評価を通し，2 つの構成要素を組み合わせることでより幅広い職業について行動が獲得できること，また，主語ベース部では他者から言及されやすい行動が多く獲得される傾向にあるのに対し，著者ベース部では対象の職業の日常に根ざした行動が多く獲得される傾向にあることを示す．

キーワード：知識獲得，職業的知識，ソーシャルメディア

Acquiring Activities of People Engaged in Certain Occupations

Miho Matsunagi[1,a)]　Ryohei Sasano[2]　Hiroya Takamura[1,3]　Manabu Okumura[1]

Received: March 8, 2018, Accepted: July 6, 2018

Abstract: We present a system to acquire the activities of people engaged in certain occupations. Our system consists of a subject-based component, which collects activities from sentences whose subjects are the target occupations, and an author-based component, which collects activities from sentences written by people engaged in the target occupations. We collect activities by these components and then acquire activities of the target occupations by calculating the χ^2 score between the target occupation and each collected activity. Through experiments, we show that we can collect activities of more diverse occupations by combining two components and also show that the subject-based component often acquires activities that are frequently mentioned by others, while the author-based component often acquires activities that are related to the target occupation and performed in the daily life.

Keywords: knowledge acquisition, occupational knowledge, social media

1. はじめに

職業の行動や実態に関する知識は様々な局面において有用であると考えられる．たとえば，将来の職業として医師を検討している人間は，医師が手術を行うことが多い職業であるとともに学術論文を読むことも多い職業だという知識を得ることで，医師になるために必要な能力を把握することが可能となる．また，嗜好や習慣について詳しい情報を知らない相手に贈り物をする際に，相手の職業を手掛かりとして適切な贈り物を考えることも可能である．このように，ある職業に特徴的な行動や実態に関する知識は就職活動や個人の属性に基づく提案など様々な場面における活用が期待できる．このような知識は，対象の職業に従事する人間への聞き取りを行うなどの方法でも構築が可能であるとは考えられる．しかしながら，多くの職業に対してこのような知識を人手で構築するコストは大きく，また，職務内容は技術革新や関係する法律の改正などによって変化する可能性があることから定期的に知識を更新する必要がある．さらに，他の職業と比較すると対象の職業に特徴的

[1] 東京工業大学
Tokyo Institute of Technology, Yokohama, Kanagawa 226–8503, Japan
[2] 名古屋大学
Nagoya University, Nagoya, Aichi 464–8601, Japan
[3] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology, Koto, Tokyo 135–0064, Japan
[a)] matsunag@lr.pi.titech.ac.jp

であると見なせる行動が存在しても，対象の職業に従事する人間はそれが特徴的な行動だと意識していない可能性も存在する．そこで，本研究では Web 上のテキストから各職業に関する行動を自動収集することにより，大規模に，かつ，他の職業と比較することで初めて特徴的であることが明らかとなるような行動も含めた知識の収集を目指す．

職業や年齢などの個人の属性に関する知識を獲得する既存研究 [1], [11] においては，知識は語単位で獲得されることが多い．このような研究では，たとえば文 (1) のような例から「映画監督」に関連する語として「映画」や「IMAX カメラ」を収集し，知識獲得に利用する．

(1)　ある映画監督は映画を IMAX カメラで撮影した。

しかし，単一の語単位では対象となる属性の詳細な特徴を獲得できない場合がある．たとえば，「映画」に関係する職業には「映画監督」や「俳優」などが存在するが，「映画監督」は「映画を撮影する」のに対し，「俳優」は「映画に出演する」と対照的な立場をとる．このような違いをとらえるために，本研究では「映画を撮影する」のような動詞の主語以外の項の 1 つとその動詞のペアを行動と定義し，対象の職業に特徴的な行動を知識として獲得する．

本研究では，対象の職業に特徴的な行動を獲得するために 2 種類の情報源を用いる．まず，1 種類目の情報源としては，文の主語が職業名となっている文を利用する．文 (1) はこのような文の例となっており，「映画監督」の行動として「映画を撮影する」および「IMAX カメラで撮影する」の 2 つの行動が抽出される．これらの文からは，行動の主体が対象の職業そのものとなっているため，高い確率で対象の職業に従事する人物による行動を収集することができると考えられる．

しかしながら，このような形式では現れにくい行動も存在する．たとえば，薬剤師は新薬の情報を得て業務で用いるために勉強会に頻繁に出席している[*1]が，Web テキスト上では薬剤師が主語である形で記述されることは稀である．これは，「薬を調合する」などの行動は一般人が薬剤師の行動として連想しやすいことから，職業名を主語とした文の形で言及されやすいのに対し，「勉強会に行く」という行動はそれらよりも職業から連想しにくく，職業名を主語とした文の形では言及されにくいからだと考えられる．

そこで，このような行動を収集するために，2 種類目の情報源として対象の職業に従事する人間によって書かれたソーシャルメディアのテキストを利用する．ソーシャルメディアでは様々な職業に従事するユーザが日常的な出来事を書き込んでいるため，1 種類目の情報源と比べ対象の職業に従事する人間の視点に立った行動が獲得できると考えられる．たとえば，薬剤師の業務に従事するユーザの中に文 (2) のようなイベントを投稿しているユーザが多く存在していれば，これらの文を情報源として利用することにより，薬剤師の行動として「勉強会に行く」という行動を収集できる．このとき，各ユーザが自身の投稿したイベントがユーザの職業に特徴的であることを意識していなくとも，多くのユーザの投稿を考慮することで他の職業に従事するユーザと比べてその行動が多くとられているかどうかを認識することが可能である．

(2)　今日は妊婦と薬剤の勉強会に行った！　久しぶりに面白い勉強会だった！

本研究では，これら 2 種類の異なる性質を持つ情報源から対象の職業に特徴的な行動の獲得を目指す．

図 1 にシステムの概要を示す．システムは 1 種類目の情報源から行動を収集する主語ベース部と 2 種類目の情報源から行動を収集する著者ベース部の 2 つの要素から構成される．主語ベース部では主語が対象の職業名である文を収集し，行動を抽出する．著者ベース部ではソーシャルメディアから対象の職業に従事するユーザを収集し，ユーザの行動を投稿から抽出する．これら 2 つの構成要素から収集した行動の対象の職業に対する特徴度合いをカイ二乗値で計算することで，最終的に対象の職業に特徴的な行動を出力する．

獲得した行動はクラウドソーシングを用いて評価する．データマイニング・知識獲得の評価では適合率と再現率の両方を考慮することが望ましいが，本研究では対象の職業に従事する人間が必ずしも意識していないような特徴的な行動を収集することが目的の 1 つであり，このような行動を網羅したデータを大規模に構築することは非常に難しいと考えられることから，適合率を中心として評価を行う．また，2 つの構成要素の特徴を明らかにするため，各構成要素により収集された行動の特徴の分析も行う．

2.　関連研究

テキストから行動を抽出する研究としては，Filatova ら [2] や Kozareva [8] の研究などが存在する．Filatova ら [2] は対象となる人物の行動を文書から抜き出したうえでその行動が対象人物の職業に特有かどうかを判別している．また，Kozareva [8] はブートストラップ法を用いて個人や組織の特徴的な行動を獲得している．しかしながら，これらの研究は「薬剤師」に対する「薬を調合する」のような典型的な行動のみを主な獲得対象としている．本研究では典型的な行動だけではなく，「薬剤師」に対する「勉強会に行く」のような対象の職業において典型的ではない行動も収集対象として考えている．

性別や年代といった人間の属性に関する知識の自動獲得を目的とした研究としては Sap らの研究 [11] や Bergsma らの研究 [1] などが存在する．Sap ら [11] は，ソーシャル

[*1] http://www.jpec.or.jp/kenshu/jyukou/kenshunintei_todoufuken.html

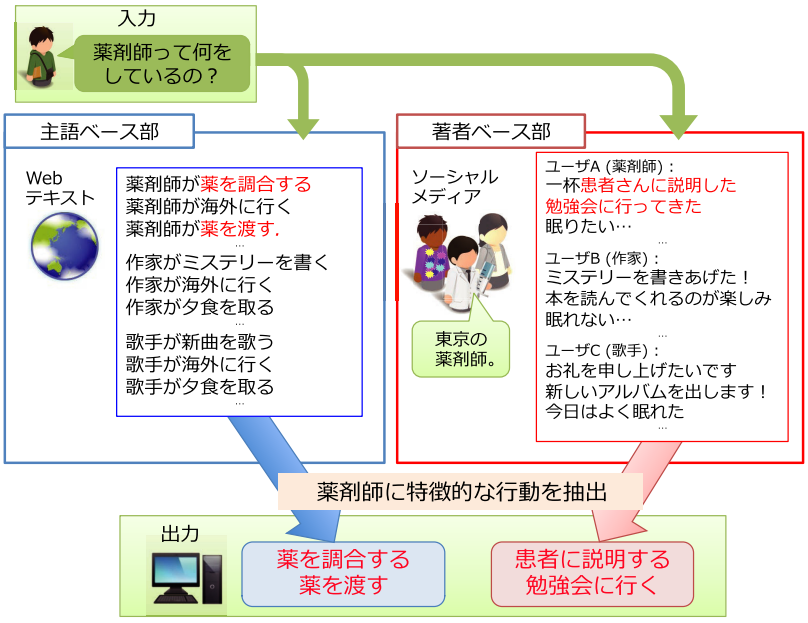

図 1　システムの概要
Fig. 1　Overview of our system.

メディア上で性別や年代などの属性を明示しているユーザの投稿から属性と関連する語の共起度合いを学習することで知識を獲得した．また，Bergsma ら [1] は Web テキストから性別を表す代名詞と共起する特定のパターンを抽出し，相互情報量を計算することでソーシャルメディア上のユーザの性別に関する知識を獲得した．しかしながら，男女に分けられる性別や 10 代から 90 代程度までに分けられる年代はクラス数が限られているため，それぞれのクラスに対してラベル付きデータを用意し教師あり学習を行うことが可能であるのに対し，職業の数は性別や年代よりも遥かに大きく，またクラス数は明確には定まっていないことから，これらの手法を職業に関連する知識の獲得にそのまま適用することは難しい．そこで，提案システムにおいて，主語ベース部は教師なしキーフレーズ抽出に類似したアプローチを採用する．

教師なしキーフレーズ抽出に関する既存のアプローチには，言語モデル [12] や共起情報のランキング手法 [4], [9], [13], [14] に基づくものなど多くのアプローチが存在するが，これらのアプローチでは基本的に対象のドメインとフレーズの共起情報からキーフレーズを抽出している．提案システムにおける主語ベース部でもこれらの研究にならい対象の職業と行動の共起情報を用いて職業に関係する行動を抽出する．

また，著者ベース部も職業と行動の共起情報を用いるが，情報源はソーシャルメディアであるため，行動の抽出方法は主語ベース部とは異なる．具体的にはユーザ自身の行動として人称が 1 人称の行動を抽出するが，本研究で扱う日本語では 1 人称代名詞は頻繁に省略されるため，各行動に対して省略されている主体を推定する必要がある．このような問題に対して Kanouchi ら [5] はソーシャルメディアで言及されている疾患およびその症状の主体を推定する分類器を構築している．本研究では職業に関する行動の収集が目的であること，また対象の職業に従事する多くのユーザによる行動を集積することで，ある程度他者の行動が含まれてしまうことによるノイズはとり除けると判断したため，教師あり学習は行わずにテキストから著者自身の行動を抽出するルールを用いることで，著者以外が主体となっている行動を取り除く．

3.　提案システム

図 1 に示すように，提案システムは主語ベース部と著者ベース部の 2 つの構成要素からそれぞれ入力された職業に特徴的な行動の候補を収集し，職業と行動の間のカイ二乗値を計算することで職業に特徴的な行動を獲得する．以下，本研究で扱う行動を定義したうえで，各構成要素について詳細を述べる．

3.1　行動の定義

本研究においては，文中の動詞とその主語を除く項 1 つの組をその動詞の主体による行動と定義する．1 つの動詞に主語以外の項が複数存在する場合は，各項についてそれぞれ独立した行動として扱う．たとえば以下の文 (3) には医者の行動として「病院で診る」および「患者を診る」が含まれていると考える．

(3)　医者が病院で患者を診る。

また，動詞には動作を表す動詞と状態を表す動詞が存在し，動作を表す動詞を収集するのが自然であるとも考えられる

表 1　カイ二乗値の算出に用いる頻度 $N_{i,j}$ の一覧．i と j はそれぞれ職業と行動を表す

Table 1　Notation for frequencies regarding for calculating the chi-square score. i and j in $N_{i,j}$ denote a target occupation and activity, respectively.

Type	Description
$N_{1,1}$	対象の職業に従事する人物が対象の行動を実行する回数
$N_{1,0}$	対象の職業に従事する人物が対象ではない行動を実行する回数
$N_{0,1}$	対象の職業に従事しない人物が対象の行動を実行する回数
$N_{0,0}$	対象の職業に従事しない人物が対象ではない行動を実行する回数

が，本研究においては「病棟に居る」が「看護師」の特徴を表すように，状態を表す動詞も職業の特徴を表すと考え，状態を表す動詞と項の組も収集対象とする．

3.2　主語ベース部

主語ベース部では，まず対象の職業が主語として現れる文から行動を抽出する．たとえば，図 1 に示すように「薬剤師が薬を調合する」と「薬剤師が海外に行く」はともに薬剤師による行動を含んでいるため，主語ベース部はこれらの行動をすべて対象の職業に特徴的な行動候補として収集する．しかしながら，収集されたすべての行動が対象の職業に特徴的な行動であるとは限らない．この例では「薬を調合する」は薬剤師の特徴的な行動だが，「海外に行く」という行動は薬剤師が他の職業より多くとるとは考えにくく，薬剤師に特徴的な行動ではないと考えられる．これらの行動候補から職業特有の行動を獲得するため，行動と職業のカイ二乗値 [10] を計算する．**表 1** にカイ二乗値（χ^2）の算出に用いる頻度 $N_{i,j}$ の一覧を示す．たとえば「薬剤師」と「薬を調合する」における $N_{0,1}$ は，「薬剤師」以外の人間が「薬を調合する」という行動をとった頻度を示す．このとき，カイ二乗値は $N_{i,j}$ を用いた次式に従って計算される．

$$E_{i,j} = \sum_{i'} N_{i',j} \sum_{j'} N_{i,j'} / \sum_{i',j'} N_{i',j'},$$

$$\chi^2 = \sum_{i,j} (N_{i,j} - E_{i,j})^2 / E_{i,j}.$$

カイ二乗値は，対象の職業における行動の出現頻度が，職業と行動が独立である場合と比べてどの程度離れているかを表しているため，対象の職業に特徴的な行動のカイ二乗値は他の行動よりも大きくなると考えられる．本研究では対象の職業と各行動におけるカイ二乗値を計算し，カイ二乗値の大きい行動をその職業に特徴的な行動として出力する．ただし，対象の職業における行動の頻度が期待値よりも小さい場合にもカイ二乗値は大きくなるため，カイ二乗値の計算は対象の職業における頻度が期待値よりも大きい行動を対象とする．

主語ベース部では大規模テキストを構文解析した解析結果から行動を抽出している．このとき，自動的に構文解析された結果には一定の割合で解析誤りが含まれるため，誤った解析結果を除外する必要がある．本研究では Kawahara ら [6] の手法に従い，対象の職業がガ格として現れており，そのガ格の係り先に曖昧性がない場合に，係り先の述語と項のペアを抽出したデータを用いている．たとえば，以下の文 (4) からは弁護士の行動として「相談を受け付ける」が抽出できるが，文 (5) では主語の「弁護士」が「受ける」と「整理する」のどちらに係るかは曖昧性が存在するため弁護士の行動としては抽出しない．

(4)　弁護士が無料で相談を受け付けます。

(5)　弁護士がネットで受けた相談を整理する。

3.3　著者ベース部

著者ベース部では，ソーシャルメディアから対象の職業に従事するユーザの行動を抽出する．ソーシャルメディアのユーザは日常的な出来事を中心に投稿を行っているため，職業に関する行動の中でも日常生活で行われる行動が多く獲得できると考えられる．このような行動を獲得するため，著者ベース部は対象の職業に従事すると考えられるユーザをソーシャルメディアから収集したうえで，ユーザ自身の行動を投稿から抽出し，職業と行動の間のカイ二乗値を計算することで最終的に職業に特徴的な行動を獲得する．

まず，対象の職業に従事すると考えられる，プロフィールに対象の職業を明記しているユーザをソーシャルメディアから収集する．ただし，すべてのユーザが記載している職業に実際に従事しているとは限らない．たとえば，ユーザの中には現在の職業ではなく将来の職業として志望している職業を記載するユーザも存在する．本研究では，プロフィールが次の制約をすべて満たすユーザをその職業に実際に従事するユーザとして収集する．

制約（A）　対象の職業名の直前の語が「元」でなく，直後が判定詞「です」，句点「。」，空白，または，行末である．

制約（B）　「父」，「母」，「姉」などの家族を表す語を含まない．

制約（C）　「夢」，「趣味」を含まない．

制約（D）　「自動」，「非公式」，「bot」，「ニュース」などの語が記載されていない．

ここで，制約（A）は「元アナウンサーです」など現在はその職業ではない可能性が高いユーザを除外する．制約（B）は「妻は医者です」など，プロフィール中の対象の職業名が家族の職業であるようなユーザを除外する．制約（C）は「夢は医者」，「趣味：探偵」など，その職業を志望しているか，趣味としてその職業に言及しているユーザを除外する．制約（D）は「医者です。自動で役立つ情報を配信

表 2　著者自身の行動を抽出するための制約一覧．これらの制約をすべて満たした行動だけが抽出される

Table 2　Rules for extracting the authors' own activities. Only activities that satisfy these constraints are extracted.

名称	詳細
主語	主語が 1 人称代名詞か省略されたものである．
目的語	目的語が 1 人称代名詞ではない．
連体修飾	行動を表す動詞は名詞を修飾するものではない．
モダリティ	文のモダリティが疑問・命令・仮定など，実際には実行されていないことを表すものではない．

します」など，ソーシャルメディア中で自動的に投稿を行う bot やスパムである可能性が高いユーザを除外する．

次に，収集されたユーザの投稿からユーザによる行動を抽出する．投稿中に記述された行動は必ずしも書き手であるユーザ自身の行動であるとは限らないため，**表 2** に示す制約によって著者自身の行動を抽出する．ただし，表 2 における「モダリティ」は行動の候補となる動詞に続く語尾の性質から判断する．これらの制約によって，以下の例においては文 (6) および文 (7) のみ下線部の行動を抽出する．

(6)　私は家で夕食を食べましたよ。

(7)　東京に着きました。

(8)　家についたら私に連絡してください。

(9)　餌を食べる犬を見た。

(10)　病院に行きなさい。

このとき，文 (8) は目的語が著者であることから，文 (9) は「食べる」が「犬」を修飾していることから，文 (10) は実際には病院には行っていないことから，それぞれ抽出されない．

そして，収集した行動に対し，主語ベース部と同様に対象の職業におけるカイ二乗値の大きい行動を職業に関係する行動として出力する．

4. 実験

4.1 データセット

行動を収集するためのデータとして，主語ベース部では Kawahara ら [7] の手順で構築した 100 億文を超える Web テキストに含まれるおよそ 65 億個の述語項構造データ[*2]から行動を抽出した．この述語項構造データは上記の 100 億文を超える Web テキストに対して 3.2 節の処理手順を適用することで構築されており，動詞ごとに Web 上に出現した述語項構造とその出現回数が記述されている．また，著者ベース部では 2013 年に日本語で投稿した Twitter ユーザの一部を Twitter API[*3]を用いて抽出した．具体的には，すべてのツイートからランダムにツイートを抽出する API[*4]を用いて 2013 年の日本語のツイートのうち約 1%を抽出し，その中に含まれるユーザから 3.3 節の手順によって対象の職業に関連付けられたユーザを抽出した．抽出した約 11,287,300 ユーザのうち，対象の職業のいずれかと関連付けられたユーザは約 32,000 人であった．

表 3　対象とした職業の一覧

Table 3　List of target occupations.

アナウンサー	作家	カメラマン	大工
コック	カウンセラー	学芸員	探偵
栄養士	医者	編集者	エンジニア
警備員	美容師	保育士	主婦
弁護士	音楽家	看護師	画家
薬剤師	駅員	パイロット	公務員
歌手	教師	劇団員	記者

また，評価実験のために次の 3 つの基準を満たす 159 職業から 28 職業をあらかじめ選び，システムへの入力とした．

(1) Wikipedia の職業リスト[*5]に職業として記載されている．

(2) 形態素解析器 JUMAN[*6] の辞書に名詞または名詞性接尾辞として登録されており，『人』というカテゴリ情報が付与されている．

(3) 前述の 65 億個の述語項構造データにおいて項として 10,000 回以上出現している．

ここで，(2) においてカテゴリ情報が『人』となっているという制約を加えているのは，職業以外の意味で使用されることが多い「質屋」や「葬儀屋」などを除外するためである．評価対象とする 28 職業を選択する際は，分散が大きくなるように「医師」と「獣医師」のような類似した職業を選択しないように留意した．また，「社長」，「部長」および「助手」は組織における位置付けを表す名称であり具体的な職業ではないこと，また「会社員」は特徴を抽出するうえで他の職業と比べあまりにも一般的であることから，評価対象には含めなかった．**表 3** に対象とした 28 職業を示す．

4.2 実験 1：ユーザの関連付け精度

予備実験として，著者ベース部において Twitter ユーザに関連付けられた職業が実際の職業とどの程度合致しているかの検証を行った．各職業につき，2 人の評価者がそれぞれ最大 100 ユーザ，合わせて最大 200 ユーザのプロフィールを確認し，関連付けられた職業に実際に従事しているかどうかを判定した．ただし，対象の職業における収集人数が 200 人未満だった場合はその職業と紐付けられた

[*2] 述語項構造データとしては京都大学黒橋・河原研究室より提供されたデータを利用した．
[*3] https://dev.twitter.com/docs
[*4] https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample
[*5] https://ja.wikipedia.org/wiki/職業一覧（2015 年 6 月閲覧）
[*6] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

表 4　対象の職業に従事するユーザ数と職業との関連付けの精度
Table 4　Populations and accuracies of collecting users engaged in target occupations.

職業	収集人数	精度（%）
アナウンサー	206	69.0
作家	3,186	79.5
カメラマン	1,664	85.5
大工	625	54.5
コック	367	66.5
カウンセラー	618	91.5
学芸員	91	89.0
探偵	9	11.1
栄養士	1,410	84.5
医者	383	59.0
編集者	1,373	96.5
エンジニア	5,843	93.5
警備員	1,528	44.5
美容師	3,527	85.5
保育士	2,819	84.5
主婦	23,556	97.0
弁護士	428	90.0
音楽家	694	90.5
看護師	2,819	90.0
画家	552	88.0
薬剤師	1,030	92.5
パイロット	290	14.0
公務員	1,502	80.0
歌手	1,348	60.0
駅員	2	50.0
教師	1,664	71.0
劇団員	239	88.5
記者	396	94.0

すべてのユーザを評価した．**表 4** の結果から，6 割以上の職業については関連付けられたユーザの 80%以上が実際に対象の職業に従事していたことが確認できた．

一方，一部の職業ではソーシャルメディア上のユーザを十分な精度または量で関連付けることができなかった．以下，関連付けの精度が低かった職業についての分析を行う．

「駅員」,「学芸員」,「探偵」については，その職業に従事するユーザをほとんど収集できなかった．これらの職業は，ソーシャルメディア上で職業を表明しているユーザ数自体が少ないため，その中でさらに 3.3 節における制約（A）から制約（D）を満たすユーザはほとんど存在しなかった．

続いて，関連付けられたユーザは多く存在したものの，精度が低かった職業としては,「歌手」,「パイロット」,「大工」,「警備員」,「医者」があげられる．「歌手」は一般人の憧れになりやすい職業であり，その職業を愛好していることを示す表現がノイズとなった．しかし，これらの表現をすべて単純なパターンで除去するのは難しい．たとえば，文 (11) では，職業名の直後が判定詞「です」となっているため,「歌手」に関連付けられたが，実際にはユーザの職業を表していない．

(11)　ジェジュンは私にとって世界一の歌手です。

また,「パイロット」は文 (12) のような特定の著名人や創作物中の登場人物を名乗るアカウントが多く存在したため，適切にユーザを収集することはできなかった．

(12)　護送船パイロット。護送中に脱走を許してしまったボラード星人を追って地球へやってきた。

「大工」や「警備員」の収集精度が低かったのは，実際には他の語であったり，他の語と複合することで意味が変化する場合が多かったためである．

(13)　鳥大工 3 年

(14)　基本ＴＬ警備員ですがフォロー返しします。

(15)　農二 2 年　サッカー部　部室警備員

たとえば文 (13) は本来「大学の工学部」を表しているが，形態素解析を行うと「鳥/大工」と分割されてしまうため誤って収集された．また，文 (14)，文 (15) のように「警備員」は複合語になると本来の警備員から意味が変化するため，実際の職業とは関係のないユーザが誤って収集された．「医者」は，人間としての医療従事者の意味以外にも，医療が提供される場所として「医院」に近い意味を持つ場合もあり,「医院」を詐称するスパム業者のプロフィールと見られる例が多く存在したため，低い収集精度となった．

このように，ソーシャルメディアを用いたユーザ収集はソーシャルメディア自体の性質や Web テキストの性質から発生したノイズの影響で十分に行えなかった職業も存在したものの，これらの職業の持つ性質が当てはまらないその他の一般的な職業であれば十分にユーザを収集することができた．

4.3　実験 2：行動の獲得精度

1 章で述べたように，本研究では各情報源から獲得された知識は互いに異なる性質を持つと仮定している．この仮定を検証するため，クラウドソーシングを用いた評価を実施した．対象の職業ごとにカイ二乗値が上位 100 件の行動をクラウドソーシングサービスのランサーズ上で作業者に提示し，職業と行動の関係性が以下の 3 段階のうちどれに該当するか尋ねた．

関係あり（明らか）：提示された行動は職業に関係する行動だと考えられ，すぐに職業に関係する行動として連想できる．

関係あり（明らかではない）：提示された行動は職業に関係する行動だと考えられ，すぐには職業に関係する行動として連想できないがよく考えれば関係すると分かる．

関係なし：提示された行動は職業に関係する行動だとは考

えられない．

ただし，実際に対象の職業に従事する人間が評価する方が本来は望ましいものの，本研究で対象とした一般的な職業については，候補となる行動を Web などで調べることで専門家以外でも評価することを可能としていること，また，すべての対象の職業に従事する人間にアンケートをとることはコスト上の問題から難しいことから，本研究では評価者を専門家に限定せず，「よく知らない職業については，インターネットや辞典などで何をしているのかを調べても構いません」と指示している．クラウドソーシング上で評価の品質を確保するため，作業者には「薬剤師」と「薬を調合する」のような，作業内容を理解していれば容易に回答できるペアを提示し，これらのペアの関係性を正答できた作業者による回答のみを評価結果として用いた．また，たとえば似たような行動であっても「小説を書く」と「文章を書く」のように異なる粒度で表現される場合があるが，収集対象とする行動の粒度を事前に明確に規定することは難しいと考えているため，作業者には上記の基準以上の細かい定義をせずに作業を依頼している．ただし，作業者ごとのタグの揺れを防ぐため，1 つのペアごとに 5 人の作業者が関係性について回答し，4.4 節の精度は 5 人中 3 人以上の作業者が「関係あり（明らか)」または「関係あり（明らかではない)」と回答した行動の数として計算されている．

1 章で述べたように，本研究では適合率に基づく評価を実施している．獲得行動数の評価を明示的には行っていないものの，実用上は 2 つの構成要素でカイ二乗値上位 100 件ずつ，合計 200 件の行動が獲得できれば十分な数になると考えている．ただし，正確な評価を行うためには十分な量の行動が獲得されている必要があるため，各手法で 200 件以上の知識が獲得された職業を評価の対象とした．その結果，28 職業中 13 職業が主語ベース部および著者ベース部における評価対象となり，11 職業が著者ベース部のみにおける評価対象となった．また，4 職業はいずれの手法においても評価対象とならなかった．

4.4　評価結果

評価結果を**表 5** に示す．以下では，まず各構成要素ごとの結果について述べる．

主語ベース部では評価を行った 13 職業のうち 10 職業では約 60%以上の精度で対象の職業に特徴的な行動を獲得していた．ただし，主婦や教師など，一部の職業については低い精度となっていた．原因としては，これらの職業には文 (16) のような広告に含まれる文章が多数含まれているが，これらの文のほとんどは職業の実態を反映していないためノイズになったことが考えられる．

(16)　主婦が FX で稼ぐ

著者ベース部においては行動の獲得精度はユーザと職業の関連付け精度と相関係数 0.48 で中程度の相関が見られた．しかしながら，一部の職業はこの傾向に該当しなかった．公務員や栄養士はユーザと職業の関連付け精度は高精度であったものの，行動の獲得精度はそれぞれ 4%，23%と非常に低い結果となった．この理由としては，これらの職業では職業に関連した書き込みをすることが少ないからだと考えられる．一方で，歌手に対するユーザと職業の関連付け精度は低い水準にあったものの，高い精度で行動を獲得することができていた．この原因としては，歌手は以下のように自身の活動を宣伝するなど，頻繁に職業に関係する投稿を行う傾向があるため，職業に従事する人間の割合が低くても職業に関する行動に言及する人間の割合は比較的高いからだと考えられる．

(17)　今日は仙北市民会館で「Soundance」という音楽とダンスのイベントに出演します♪

(18)　今日は埼玉商工会議所のパーティで歌います。

次に，両方の構成要素で評価対象になった 13 職業に着目し，構成要素の性能比較を行った．**図 2** は 13 職業における 2 つの構成要素の精度を散布図で比較している．図 2

表 5　構成要素ごとの精度．参考のため，4 列目に表 4 の著者ベース部におけるユーザの職業への関連付け精度を示す

Table 5　Scores of each component. We also show in the rightmost column the accuracies of collecting users by the author-based component in Table 4.

職業	主語ベース部	著者ベース部	職業の関連付け精度
アナウンサー	66%	52%	69.0%
作家	70%	72%	79.5%
カメラマン	82%	95%	85.5%
コック	90%	33%	66.5%
カウンセラー	62%	17%	91.5%
医者	64%	10%	59.0%
エンジニア	59%	59%	93.5%
主婦	16%	78%	97.0%
弁護士	52%	33%	90.0%
看護師	69%	79%	90.0%
歌手	82%	78%	60.0%
教師	41%	54%	71.0%
記者	59%	34%	94.0%
大工	-	10%	54.5%
栄養士	-	23%	84.5%
編集者	-	73%	96.5%
警備員	-	1%	44.5%
美容師	-	48%	85.5%
保育士	-	67%	84.5%
音楽家	-	85%	90.5%
画家	-	85%	88.0%
薬剤師	-	62%	92.5%
公務員	-	4%	80.0%
劇団員	-	47%	88.5%
平均	62.5%	50.4%	-

表 6　正しく獲得できた行動の内訳．$N_{obvious}$，$N_{non\text{-}obvious}$，N_{other} はそれぞれ A，B，C の行動の種類数，また N_{all} はこれらの合計を示す

Table 6　The characteristics of correctly acquired activities. $N_{obvious}$, $N_{non\text{-}obvious}$, and N_{other} denote A, B, C respectively and N_{all} denotes the sum of them.

構成要素	N_{all}	$N_{obvious}$	$N_{non\text{-}obvious}$	N_{other}	$N_{non\text{-}obvious}$ / ($N_{obvious}$ + $N_{non\text{-}obvious}$)
主語ベース	812	516	279	17	35.1%
著者ベース	694	369	299	26	**44.8%**

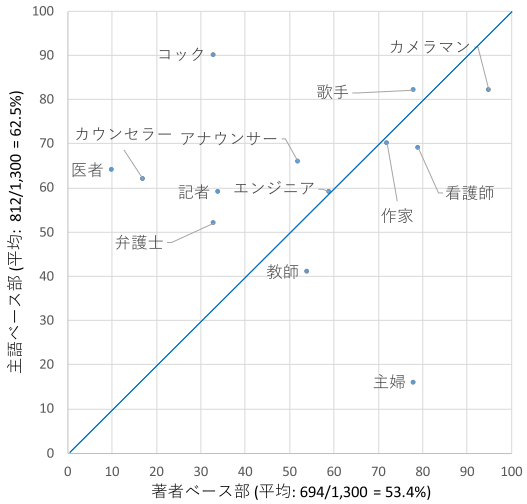

図 2　2 つの構成要素の比較．x 軸と y 軸はそれぞれ著者ベース部と主語ベース部の精度を表す，参考のため，軸ラベルに 13 職業における各構成要素の平均精度を示す

Fig. 2　Comparison of activities acquired by the two components. The x and y axis denote the score of a author-based component and a subject-based component, respectively. For reference, we also show the average accuracies of each component in axis titles.

から，職業によっては 2 つの構成要素の性能には大きな差が存在することが分かる．具体的には，主語ベース部は著者ベース部よりも医者やコックについては高い精度で行動を収集できているが，主婦や教師については著者ベース部の方が主語ベース部よりも高い精度となっている．このことから，ある程度高い精度で幅広い職業の行動を獲得するためには，これらの 2 つの高精度になる部分を組み合わせる必要があるといえる．また，正しく獲得できた行動の数は，13 職業の中で最も少なかった医者でも 2 つの構成要素で合わせて 74 個の行動を獲得できており，実際に人間に提示する行動の数としては十分だと考えている．

ただし，獲得した行動の中には文 (19) のように単独で解釈するには不自然な行動も存在した．

(19)　学会で行く（医者）

このような行動を獲得した原因としては，本研究では動詞とその項 1 つを行動として扱っていることから，本来は「学会でオーストラリアに行く」のように複数の項が必要な文であったにもかかわらずその項が 1 つずつ独立した行動として提示されたことが考えられる．

4.5　分析

続いて，2 つの構成要素によって正しく獲得できた行動の性質について更に分析を行った．4.1 節で述べたように，クラウドソーシングを用いた評価では，対象の職業と行動のペアの関係性が「関係あり（明らか）」か「関係あり（明らかではない）」のどちらかだと 5 人中 3 人以上の作業者が判定した場合，その行動は対象の職業に関係する行動だと判断した．これらの対象の職業に関係する行動だと判定された行動を，評価の詳細によってさらに 3 つに分類した．

A（obvious）：「関係あり（明らか）」と判断した作業者の数が「関係あり（明らかではない）」と判断した作業者の数より大きい．

B（non-obvious）：「関係あり（明らかではない）」と判断した作業者の数が「関係あり（明らか）」と判断した作業者の数より大きい．

C（other）：「関係あり（明らか）」と判断した作業者の数が「関係あり（明らかではない）」と判断した作業者の数と等しい．

表 6 に内訳を示す．ここで，N_{all} は正答と判定された行動の総種類数，また，$N_{obvious}$，$N_{non\text{-}obvious}$，N_{other} はそれぞれ A，B，C に分類された行動の種類数を表す．表 6 から，著者ベース部は主語ベース部よりも多くの「関係あり（明らかではない）」と判断されやすい行動を獲得できていることが分かる[*7]．この差をフィッシャーの正確確率検定 [3] によって検定したところ，有意水準 0.01 で有意であった．この結果から，2 つの情報源からは異なる行動が獲得できるという予想が裏付けられた．

表 7 に出力例を示す．著者ベース部は対象の職業で日常的に行われる，職業の実態に即した行動を獲得している．たとえば，弁護士の行動としては「書面を起案する」という行動が収集されているが，弁護士における「書面」は民事訴訟で自身の主張を記述する「準備書面」を指す法律用語であり，準備書面を依頼者のために用意するという弁護士の実態に即した行動を獲得できている．また，エンジニアの行動として収集された「勉強会に参加する」や「資料を作る」のように，すぐには思いつかないが実際には行われていると考えられる行動も一定数収集できている．しか

[*7] このとき，「関係あり（明らかでない）」の割合を 13 職業のマクロ平均で計算を行うと，正しく獲得できた行動数が少ない職業における結果は信頼性が低いにもかかわらず全体の結果に影響してしまうため，ミクロ平均によって割合を計算している．

表 7　それぞれの構成要素によって獲得された行動の例．行動の後の括弧内の数字はクラウドソーシングの評価の「明らか」,「明らかではない」,「無関係」を順に示す

Table 7　Examples of activities acquired in each component. Numbers in brackets denote the number of crowd-workers who evaluated the activity as (obvious, non-obvious, irrelevant), respectively.

職業	主語ベース部	著者ベース部	職業	主語ベース部	著者ベース部
弁護士	相談を受け付ける (1,4,0)	書面を起案する (1,2,2)	エンジニア	プログラムに集中する (4,1,0)	資料を作る (0,3,2)
	弁護に当たる (5,0,0)	弁護士を募集する (3,2,0)		トラブルに対応する (0,4,1)	勉強会に参加する (0,3,2)
	弁護団に加わる (5,0,0)	事件を受ける (2,3,0)		修理に伺う (4,1,0)	コードを書く (3,2,0)
医者	患者を騙す (1,4,0)	学会で行く (0,4,1)	カメラマン	スタジオで撮影する (5,0,0)	写真展に行く (0,5,0)
	手術に挑む (5,0,0)	下を向く (0,0,5)		フラッシュを浴びせる (4,0,1)	カメラを守る (4,1,0)
	眼鏡を勧める (0,3,2)	税金を払う (0,0,5)		モデルを撮影する (5,0,0)	写真をレタッチする (2,3,0)
主婦	ビジネスで稼ぐ (0,0,5)	洗濯物を干す (4,1,0)	看護師	身体を拭く (1,4,0)	夜勤から帰る (1,4,0)
	起業を選ぶ (0,1,4)	弁当を作る (5,0,0)		脈拍を測定する (4,1,0)	患者を受け持つ (4,1,0)
	節約術を紹介する (2,3,0)	娘を連れる (3,2,0)		点滴に来る (4,1,0)	病棟で食べる (2,2,1)

し，対象の職業に限定されない行動を獲得してしまうようなケースも確認された．たとえば，医者の行動として，著者ベース部によって「税金を払う」という行動が獲得されているが，医者がこの行動を他の職業よりも多く行うとは考えにくく，医者に特徴的な行動であるとはいえない．これらの特徴的ではない行動が獲得される原因としては，ソーシャルメディアに日常的な行動をどれだけ投稿するかは職業ごとに異なり，日常的な投稿が多い職業では誰でも行う行動の割合が他の職業よりも大きくなることから特徴的だと見なされてしまったことが考えられる．

一方，主語ベース部は他者がその職業の典型的行動として連想するような行動を多く獲得している．たとえば，弁護士に対しては「弁護に当たる」，また看護師では「脈拍を測定する」のような対象の職業における典型的な行動が獲得されている．しかし，対象の職業で実際には行われていないような奇妙な行動を獲得してしまうようなケースも存在した．たとえば，医者では「患者を騙す」という行動が収集されているが，この行動は一般的な医者の行動であるとはいえない．これらの実態を反映していない行動は特異な例として書名やニュースの見出しに出現することから，主語ベース部によって獲得されてしまったと考えられる．

また，行動と職業の関係度合いに影響する要因としては，対象の職業に従事する人間集合の中でその行動を行っている人間の割合などもある程度考えられる．表 7 において，主婦の行動として収集された「ビジネスで稼ぐ」は作業者全員が「関係なし」と判断し，また医者の行動として収集された「眼鏡を勧める」は他の行動と比べて「関係なし」と判断される割合が高くなっている．この原因としては，これらの行動は対象の職業に従事する人間の一部しか行っていないとみなされることが考えられる．ただし，医者の「手術に挑む」については，すべての医者が手術を行うわけではないにもかかわらず全員が「関係あり（明らか)」と回答している．この原因としては，少なくとも「手術に挑む」ことが可能なのはほぼ医者に限られているということが医者の中で「手術に挑む」という行動をとる医者の割合よりも判断に影響を与えたことが考えられる．

実際に収集した行動を確認すると，提案手法により獲得された行動はある程度の多様性を持っていることが分かる．たとえばエンジニアについては，「開発をサポートする」「修理に伺う」「プロジェクトに従事する」のように，顧客からシステムを受注してプロジェクトに従事したりシステムの保守を行ったりする一方で「勉強会に行く」,「ドキュメントを読む」,「本を買う」，また「資格に挑戦する」といった自己研鑽に努めることもあるという複数の側面からの行動に加え，「終電で帰る」などの生活習慣が分かるような行動も獲得している[*8]．また，作家について獲得した行動には「漫画を書く」など純粋な執筆活動を行う一方で，「個展を開催する」「作品を展示する」などのアウトプットに関わる面や「描き方を解説する」といった指導に関わる面も存在し，作家の実態について様々な側面からの知識が得られている．定量的な有用性や網羅性の評価については今後の課題となるものの，職業に従事する人間がその生活において行う様々な行動について実態を知るうえでの有用性は存在すると考えられる．

最後に提案手法の適用範囲について考察する．本研究の評価は，Web 上で高頻度に出現した 28 職業を対象に実施したが，評価対象とする職業を選定する際に使用した 4.1 節の 3 条件に適合する職業は，「社長」などのような役職を表すものを除いても他に 100 職業以上存在しており，これらの職業については実験に適用された職業と同等程度の性能で行動を獲得できると考えられる．また，実験で使用した 65 億個の述語項構造データでは長い複合名詞は項として収集されていなかったことから，たとえば「放射線技師」

[*8] ただし，エンジニアは本研究の収集対象であった Web と非常に親和性の高い IT 系エンジニアに分布が大きく偏っていたことから，獲得できた行動はほぼ IT 系エンジニアの行動に限定されていた．

などの職業は上記の 3 条件に適合する職業には含まれないものの，日本には 50,000 人以上の放射線技師が存在しており[*9]，Web 上にも相応数の放射線技師に関する情報が存在することが期待できる．したがって，より広範囲の職業に対しても同程度の精度で行動を獲得することが可能であると期待できる．

5. おわりに

本論文では，与えられた職業に特徴的な行動を獲得するためのシステムを提案した．提案システムは，対象の職業名が主語として現れる文から行動を抽出する主語ベース部と，ソーシャルメディアにおいて対象の職業に従事するユーザの投稿から行動を抽出する著者ベース部の 2 つの要素から構成されている．28 職業を対象に行動の獲得実験を行い，獲得した行動の精度をクラウドソーシングを用いて評価した結果，主語ベース部で獲得された行動の精度は平均 62.5%，著者ベース部で獲得された行動の精度は平均 51.4%の精度であった．また，獲得された行動の性質を分析することで，主語ベース部では他者がその職業の典型的行動として連想するような行動が多く獲得されるのに対し，著者ベース部では典型的行動として連想するような行動ではないものの対象の職業において日常的に行われる特徴的な行動が多く獲得される傾向があることを示した．今後の課題としては，2 つの構成要素を密に統合することでより頑健なシステムの構築を行うことが考えられる．

参考文献

[1] Bergsma, S. and Van Durme, B.: Using Conceptual Class Attributes to Characterize Social Media Users, *Proc. 51st Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.710–720 (2013).

[2] Filatova, E. and Prager, J.: Tell Me What You Do and I'll Tell You What You Are: Learning Occupation-related Activities for Biographies, *Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (*HLT/EMNLP*), pp.113–120 (2005).

[3] Fisher, R.A.: On the interpretation of χ^2 from contingency tables, and the calculation of P, *Journal of the Royal Statistical Society*, Vol.85, No.1, pp.87–94 (1922).

[4] Florescu, C. and Caragea, C.: PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents, *Proc. 55th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.1105–1115 (2017).

[5] Kanouchi, S., Komachi, M., Okazaki, N., Aramaki, E. and Ishikawa, H.: Who caught a cold? – Identifying the Subject of a Symptom, *Proc. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (*ACL-IJCNLP*), pp.1660–1670 (2015).

[6] Kawahara, D. and Kurohashi, S.: Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis, *Proc. 19th International Conference on Computational linguistics* (*COLING*), pp.425–431 (2002).

[7] Kawahara, D. and Kurohashi, S.: Case Frame Compilation from the Web using High-Performance Computing, *Proc. 5th International Conference on Language Resources and Evaluation* (*LREC*), pp.1344–1347 (2006).

[8] Kozareva, Z.: Learning Verbs on the Fly, *Proc. 24th International Conference on Computational Linguistics* (*COLING*), pp.599–610 (2012).

[9] Mihalcea, R. and Tarau, P.: TextRank: Bringing Order into Texts, *Proc. 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.404–411 (2004).

[10] Miller, R. and Siegmund, D.: Maximally Selected Chi Square Statistics, *Biometrics*, Vol.38, No.4, pp.1011–1016 (1982).

[11] Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L. and Schwartz, H.A.: Developing Age and Gender Predictive Lexica over Social Media, *Proc. 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.1146–1151 (2014).

[12] Tomokiyo, T. and Hurst, M.: A Language Model Approach to Keyphrase Extraction, *Proc. ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pp.33–40 (2003).

[13] Wan, X. and Xiao, J.: Single Document Keyphrase Extraction Using Neighborhood Knowledge, *Proc. 23rd AAAI Conference on Artificial Intelligence* (*AAAI*), pp.855–860 (2008).

[14] Zha, H.: Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering, *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR*), pp.113–120 (2002).

[*9] http://www.mhlw.go.jp/file/05-Shingikai-10801000-Iseikyoku-Soumuka/0000200803.pdf

馬緤 美穂

2013 年東京大学教養学部超域文化科学科卒業．2015 年東京工業大学総合理工学研究科博士前期課程修了．同年同研究科博士後期課程に進学．自然言語処理を専門とし，特にソーシャルメディアの活用に興味を持つ．

笹野 遼平　（正会員）

2009 年東京大学大学院情報理工学系研究科博士課程修了．京都大学特定研究員，東京工業大学助教を経て，2017 年より名古屋大学准教授．博士（情報理工学）．自然言語処理，特に照応解析，述語項構造解析に関する研究に従事．言語処理学会，人工知能学会各会員．

高村 大也　（正会員）

1997 年東京大学工学部計数工学科卒業．2000 年同大学大学院工学系研究科計数工学専攻修了（1999 年はオーストリアウィーン工科大学で研究）．2003 年奈良先端科学技術大学院大学情報科学研究科博士課程修了．博士（工学）．2003 年東京工業大学助手，のち助教，准教授を経て，2017 年同教授．2017 年より産業技術総合研究所人工知能研究センター研究チーム長を兼任．計算言語学，自然言語処理を専門とし，特に機械学習の応用に興味を持つ．言語処理学会，ACL 各会員．

奥村 学　（正会員）

1989 年東京工業大学大学院理工学研究科博士課程修了．同年東京工業大学助手．1992 年北陸先端科学技術大学院大学助教授，2000 年東京工業大学助教授，2009 年同教授，現在に至る．工学博士．自然言語処理，知的情報提示技術，語学学習支援，テキスト評価分析，テキストマイニングに関する研究に従事．電子情報通信学会，人工知能学会，AAAI，言語処理学会，ACL，認知科学会，計量国語学会各会員．

（担当編集委員　戸田 浩之）