機械学習を適用した追加情報収集支援システムの実装と評価

髙橋和子[1]　奥村学[2]　鈴木泰山[3]　鈴木佑京[3]
[1]敬愛大学国際学部　[2]東工大科学技術創成研究院　[3]（株）ピコラボ
takak@u-keiai.ac.jp　oku@lr.pi.titech.ac.jp　{taizan, ukyo}@picolab.jp

概要

本研究では，自由回答で収集したデータを調査完了後にカテゴリに変換する作業（コーディング）の支援として，回答に情報が不足するか否かを調査の現場で判断し，必要な場合は，選択肢を提示することで回答者により情報を追加できるシステムの構築を行ってきた．今回，社会調査では必須の「職業コーディング」を対象とするシステムの実装を終え，実際にコーダによる評価を行った結果，コーダのみならず自動コーディングにおいても有効性が確認できた．本稿ではこの結果について報告する．

1 はじめに

社会調査においては，調査の目的により，データを自由回答で収集し，調査完了後にカテゴリに変換する作業（コーディング）が必要で，その代表例は「職業コーディング」（自由回答を含むデータで収集される職業情報に約 200 種類あるコードのいずれか 1 つを付与する作業）である[1][2]．ここで，職業情報とは，従業先事業の種類（自由回答），仕事の内容(自由回答)，地位（選択回答），役職（選択回答），従業先の規模（選択回答）である．

コーディングは人手（コーダ）で行われるが，大規模調査ではコーダの作業量と負担は膨大である．職業コーディングの場合は，これをバッチ処理により支援する自動コーディングシステムが開発されて以来, SSM 調査（Social Stratification and Social Mobility Survey）や JGSS（日本版 General Social Surveys）等の大規模調査で多用されている[3].

しかし，回答に含まれる情報が曖昧であったり不十分な場合は，コーダのストレスは軽減されず，またコーダ，自動コーディングのいずれにおいても正しいコードが付与されない可能性が高い．回答に情報が不足する場合は，回答者自身にその場で適切な情報を追加してもらうことが有効であるが，調査員も回答者もコードの定義を熟知していないため，この判断を行うのは困難である．

そこで，これをコンピュータに行わせ，必要な場合に回答者から情報を追加してもらう方法として，職業コーディングを対象に「調査現場における追加情報収集支援システム」の構築を進めてきた[4, 5]．提案システムは，調査員がタブレットで利用することを想定した Web システムで，情報が不足すると判断した場合に追加のキーワード選択を促す仕組みをもつ．現在，システムは実装を終え，システムが提示した情報の中から選択された語を初期回答に追加した回答を得ることが可能となったため，この回答と最初の回答（以下，初期回答と呼ぶ）に対するコーダと自動コーディングによるコーディングを実施し，評価を行った．この結果について本稿で報告する．

2 方法（アルゴリズム）

提案システムのアルゴリズムを示す．

STEP1　データ入力とサーバへの送信
STEP2　自動コーディング
STEP3　情報不足の判定（情報不足と判定されなかった場合は STEP4 に進まず終了）
STEP4　追加情報の提示と収集（収集した情報を初期回答に追加し, STEP2 に戻る）

2.1 データ入力とサーバへの送信

調査員は調査現場で回答者から得られた回答をタブレットに入力し，システムが置かれたクラウドサーバに送信する．

2.2 自動コーディング

提案システムでは前述した自動コーディングシステムを利用するが, Web で入力された情報に基づきオンラインで処理を行う．ルールベース手法とサポートベクターマシン（SVM）を組み合わせた方法で，one-versus-rest 法により多値分類に拡張されている．素性は，従業先の規模を除く職業情報，学歴（選択回答），SVM に先だって実行するルールベース手法に

より出力されたコードである．自由回答は品詞付き形態素を素性番号に変換するが，同じ形態素であっても，「従業先事業の種類」と「仕事の内容」のどちらに出現したかにより素性番号が異なる．

自動コーディングでは，第 1 位に予測された結果に対し，複数の分類スコア（分離平面からの距離）を利用し,最も高いレベル A から最も低いレベル E まで 5 段階の「確信度」を付与する．ここで, rank1, rank2 は，それぞれ SVM により第 1 位，第 2 位に予測されたコードに付随して出力される分類スコアを示す．また，α , β は閾値で, 2005 年 SSM 調査データセット（12,500 事例）を用いた実験により，α =3，β =0.4 とした．

A：rank1>=0 かつ rank2<0, rank1-rank2>= α

B：rank1>=0 かつ rank2<0, rank1-rank2< α

C：rank1>=0 かつ rank2>=0

D：rank1<0 かつ rank2<0, rank1-rank2>= β

E：rank1<0 かつ rank2<0, rank1-rank2< β

確信度ごとの正解率（全事例中，正解であった事例の占める割合）と出現率（全事例の中で 該当する事例が出現した割合）は，自動コーディングシステムが処理した 4 種類の国内・国際標準の職業・産業コーディングの実験の結果，コーディングの種類やコードの数（約 60 個～400 個）が異なっていても，安定して「レベル A（約 95%）>>レベル B（約 70～90%）>>レベル C>レベル D>>レベル E（約 20%～35%）」の順で，レベル E は際だって低かった[4].

2.3 情報不足の判定

「回答が情報不足」の場合は正解率が低いと考え，職業コーディングにおいては，「確信度が最も低いレベル」を「確信度レベル E」に限定し，これが付与された場合を情報不足であると判定する．

2.4 追加情報の提示と収集

追加情報の提示は，次の手順による．

あらかじめ 過去の事例から， 自動コーディングシステムが予測した結果が不正解であった場合の正解コードを調査し，混同されやすいコード対の情報として「不正解コードー正解コード対応表」を作成しておく．またこの対応表をもとに，混同しやすい職業を記述した説明文の事例から，両者を弁別するのに有効な語を抽出した「不正解コードー正解コード弁別語表」も作成しておく．

自動コーディングシステムにより第 1 位に予測されたコードが不正解であると想定し，これをキーに「不正解コードー正解コード対応表」を検索し，対応する正解コードを見つける．次に，正解コードと不正解コードのペアにより「不正解コードー正解コード弁別語表」を検索し，該当するペアの弁別語を選択肢として提示する[i].

図 1 は，初期回答が自動コーディングにより，コード「628」（鋳物工、鍛造工、金属材料生成作業者）と予測された場合，「不正解コードー正解コード対応表」により，対応するコードとして，「633」（一般機械器具組立工・修理工），「677」（電気工事・電話工事作業者），「503」（機械・電気・化学技術者）を見つけ，「不正解コードー正解コード弁別語表」により，「628」とこれらのコードとの弁別語である 「一般機械」，「電気工事」と「電話工事」，「電気技術者」を選択肢として提示した例である．

今回選択肢として提示する語は, [4]により提案された 4 種類のうち最も有効であった方法を用いたが，提示方法の変更は容易である．

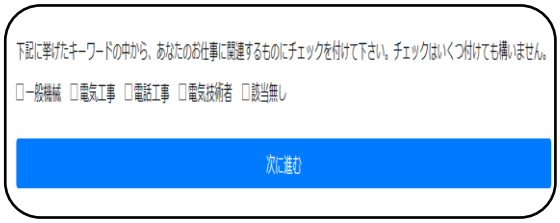


図 1　選択肢提示の例（コード「628」の場合）

この選択肢の中から回答者に 1 個以上を選んでもらい，初期回答に追加する．その際，適切な選択肢がない場合またはある場合でも，追加したい情報があれば，次画面で自由回答を入力できる．

STEP4 で収集された語を初期回答に追加して STEP2 に戻り，再度自動コーディングを行う．今回の実装では, STEP2 への戻りを 1 回のみとしたが，繰り返しの条件について検討する必要がある．

3 実験

今回は対面でデータ収集を行っていないため，提案システムのうち情報不足と判断した際に追加のキーワード入力を求める部分の有効性を調査する．調

[i] 自由回答ではなく選択肢とした理由は，回答空間を大きくしないためである．

査は情報を追加した回答と初期回答に対するコーディングの状況の比較により行なう． 両者の比較は，コーダだけでなく自動コーディングについても行う．

コーダはC大学「社会調査法」受講生6名（3年生）で，事前にコードの定義マニュアル[6]を与え，1時間のレクチャーを行ったが，大規模調査でコーダを務める大学院生や研究者ほどには熟練していない．

今回の実験ではすでに正解が分かっているため，回答者が追加すると思われる情報を選択または入力したため，実際より正解率が高くなる可能性がある．

3.1 実験設定

実験では，訓練事例を 2000 年～2006 年までのJGSS データセット（計 33,711 事例），評価事例をJGSS-2008 データセットのうち，確信度レベルが E と判定された 201 事例（全体の 7.6%）とした．

以下では，初期回答を A，情報を追加した回答をB とよぶ．情報の追加は，事例ごとに正解コードを参照し，該当する選択肢が提示された場合はこれをチェックし，ない場合は自由回答を入力する方法で行った．評価事例は，すべての事例についてデータセットA とデータセットB の 2 種類を用いた．

データセットA，データセットB をそれぞれ 6 つに分割し，サンプル番号順に A1～A6, B1～B6 と名付け，各々のデータセットについて 1 名がコーディングを担当した．作業の時間帯を前半と後半に分け，前半は 3 名が A1～A3．別の 3 名が B1～B3，後半は前半で A を担当した 3 名が B4～B6, B を担当した 3 名が A4～A6 を担当した．この作業をデータセットを変えて 2 回行ったため，どのデータセットも異なる 2 名がコーディングを行ったことになる．

コーダには，事例ごとに，「職業コード」と「ストレス度（「1：ほとんどなし（低）」「2：やや感じた（中）」「3：非常に感じた（高）」）」，また「1つのデータセットを処理した開始時刻と終了時刻」の記入を依頼した．

評価尺度は，コーダについては，正解率，ストレス度，処理時間（分）を用い，1 回目と 2 回目の平均値で評価する．ただし，今回のコーダは熟練していないために，回を重ねる効果が生じる可能性もあり，1 回目と 2 回目の比較も行う．

自動コーディングについては，正解率，提示機能の有効性を評価する．提示機能の評価は，Bにおける正解が，システムが提示した選択肢によるものか，自由回答によるものかを調査する．

3.2 実験結果

提案システムは，データセットA に対する自動コーディングの結果，各事例に対して平均 2.7 個（最大 7 個，最小 2 個）の選択肢を提示できた[ii]．

3.1.1 コーダの正解率

コーダの正解率は，表 1 に示すように，データセットB では平均 52.7%（1 回目 53.7%，2 回目 51.7%）で，データセットA より平均 16.7 ポイント（1 回目 18.4 ポイント，2 回目 14.9 ポイント）向上した．1 回目と比較すると，2 回目は平均 3.9 ポイント向上しているが，データセットA は 1.5 ポイント向上， データセットB は 2.0 ポイント低下しているため，2 回目の効果は考えられず，正解率向上におけるデータセットB の有効性が認められる．

表 1　コーダの正解状況（平均）　単位：事例数

	B 正解	B 不正解	計	**A 正解率**
A 正解	51.0	21.5	72.5	0.361
A 不正解	55.0	73.5	128.5	
計	116.0	85.0	201	
B 正解率	0.527			

3.2.2 コーダのストレス度

コーダのストレス度は，コーダ自身による評価のために個人差があるが，表 2 に示すように，データセットB で 1 つの事例につき平均 0.25（1 回目 0.30，2 回目 0.19）低下した．1 回目と比較すると，2 回目は平均 0.11/事例低下したが（データセットA は 0.13/事例，データセットB は 0.02/事例），データセット間の違いほど大きくないため，ストレス低下におけるデータセットB の有効性は認められる．

表 2　コーダのストレス度（平均）　単位：人数

ストレス度	A	B	**差（B-A）**
1：低: （割合）	72.0 （35.8%）	104.0 （51.7%）	32.0 **15.9 ポイント**
2：中 （割合）	76.5 （38.1%）	61.5 （30.6%）	-15.0 **-7.5 ポイント**
3：高 （割合）	52.5 （26.1%）	35.5 （17.7%）	-17.0 **-8.4 ポイント**
平均	1.91	1.66	-0.25

[ii] 回答者の負担を考慮し，選択肢は最大 7 個に限定した．またどの場合も「該当無し」を提示した．

3.2.3 コーダの処理時間

コーダの処理時間は，データセット B で平均 0.3 分/事例（1 回目 0.4 分，2 回目 0.1 分）短縮された．1 回目と比較すると，2 回目は平均 0.2 分/事例ほど短縮され（データセット A は 1.1 分/事例，データセット B は 0.8 分/事例），2 回目の効果とデータセットによる効果の違いは 0.1 分（6 秒）程度である．職業コードは 1 つに絞る必要があるため，情報が増えたことで判断に迷った可能性が考えられる．

3.2.4 自動コーディングの正解率

システムの正解率は，表 3 に示すように，データセット A は 49.3%，データセット B は 58.7%で，いずれもコーダより高いが，データセット B で向上した程度は 9.5 ポイントでコーダの方が高い．

表 3　システムの正解状況　　単位：事例数

	B 正解	B 不正解	計	**A 正解率**
A 正解	96	3	99	**0.493**
A 不正解	22	80	102	
計	118	83	201	
B 正解率	**0.587**			

3.2.5 提案システムの提示機能

コーダ，自動コーディングのいずれにおいてもデータセット B で正解率が向上したが，これには回答者が追加した自由回答による効果もあるため，提案システムが適切な選択肢を提示したかどうかを評価するには，データセット B で追加された情報が選択肢であるのか否かを調査する必要がある．データセット B では，選択肢のみが 97 事例（うち 8 事例は選択肢のみを 2 個追加），自由回答のみが 95 事例，選択肢と自由回答の両方が 7 事例，どちらも追加なし（初期回答と変わらず）が 2 事例で，選択肢と自由回答はほぼ同数であった．

この状況と，自動コーディングとコーダの場合における正解状況との関連を報告する．まず，自動コーディングにおいては，データセット B で正解であった 118 事例（表 3）に追加された情報は，表 4 に示すように，選択肢のみが 51 事例（43.2%），自由回答のみが 64 事例（54.2%），両方が 3 事例（2.5%）で，選択肢より自由回答の方がやや多い．

表 4 に追加された情報が選択肢のみか自由回答のみか両方かを，2 つのデータセットにおける正解／不正解の状況ごとに示す．同様に，コーダにおける状況を表 5 に示す．

表 4　自動コーディングが正解した場合の追加情報

	選択肢	自由回答	両方
A 正解 B 正解	63	27	6
A 不正解 B 正解	21	1	0
B 正解	**84**	**28**	**6**
A 正解 B 不正解	1	1	1
A 不正解 B 不正解	12	66	0

表 5　コーダが正解した場合の追加情報（平均）

	選択肢	自由回答	両方
A 正解 B 正解	25.5	21.0	4.5
A 不正解 B 正解	27.0	27.0	1.0
B 正解	**52.5**	**48.0**	**5.5**
A 正解 B 不正解	10.5	9.5	1.5
A 不正解 B 不正解	34.0	37.5	0

データセット B に追加する情報を選択肢に限定した場合の正解率は，提案システムは 41.8%であるが，コーダは平均 29.1%（1 回目と 2 回目がほぼ同じ値）と大きく低下している．該当する選択肢がない場合は自由回答で補うことができ，その有効性も示されたが，今後の課題として，選択肢の提示機能を向上させる改善を行う必要がある．

4 おわりに

本稿では．社会調査において収集される自由回答に分類に必要な情報が含まれているか否かについて，職業コーディングを対象に，機械学習を適用した自動コーディングにより調査現場で判断し，不足すると判定した場合は，候補の情報をその場で回答者に提示して追加してもらうシステムを提案した．提案システムは，コーダにおいても自動コーディングにおいても有効性を示した．

今後の課題は次の 3 つである．1 つ目は，提案システムの改善で，例えば，追加情報収集の終了条件の設定や選択肢提示機能の向上を行う必要がある．2 つ目は，提案システムに対する調査員や回答者による評価である．しかし，昨今の状況から，調査員が回答者と対面する現地調査が困難となり，これに代わるものとして，オンライン調査の研究が進み[7]，以前と違って学術調査においても多く利用されるようになってきた．提案システムは，オンライン調査への対応が容易であると考えられるため，3 つ目としてこの拡張も検討したい．

謝辞

2005 年 SSM 調査データの利用に関して，2005 年 SSM 調査研究会の許可を得た.
日本版 General Social Surveys（JGSS）は，大阪商業大学 JGSS 研究センター（文部科学大臣認定日本版総合的社会調査共同研究拠点）が，東京大学社会科学研究所の協力を受けて実施している研究プロジェクトである.

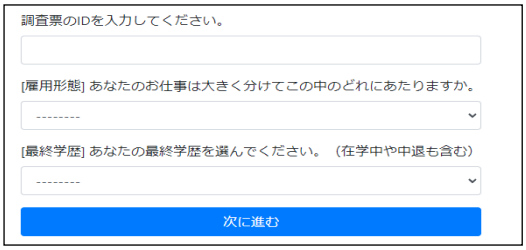
参考文献

1. **原純輔，海野道郎.** 社会調査演習. 東京大学出版会, 1984.
2. **轟亮，杉野勇.** 入門・社会調査法. 法律文化社, 2010.
3. 社会学における職業・産業コーディング自動化システムの活用. **高橋和子，多喜弘文，田辺俊介，李偉.** 自然言語処理, 2017 年, 第 24 巻第 1 号. 135-170.
4. 機械学習を適用した自由回答収集時における有効情報追加システムの構想ー職業コーディングを例としてー. **高橋和子.** データ分析の理論と応用, 2018 年, 第 7 巻第 1 号. 21-42.
5. 機械学習の適用による社会調査現場での追加情報収集支援システム, 高橋和子, 奥村学, 鈴木泰山, 清家大嗣. 言語処理学会第 26 回年次大会論文集, 2020.
6. 1995 年 SSM 調査研究会. SSM 産業分類・職業分類（95 年版）修正版. 1995 年 SSM 調査研究会, 2006.
7. **大隈昇，鳰真紀子，井田潤治，小野裕亮.** ウェブ調査の科学　The Science of Web Surveys 調査計画から分析まで. 朝倉書店, 2019.

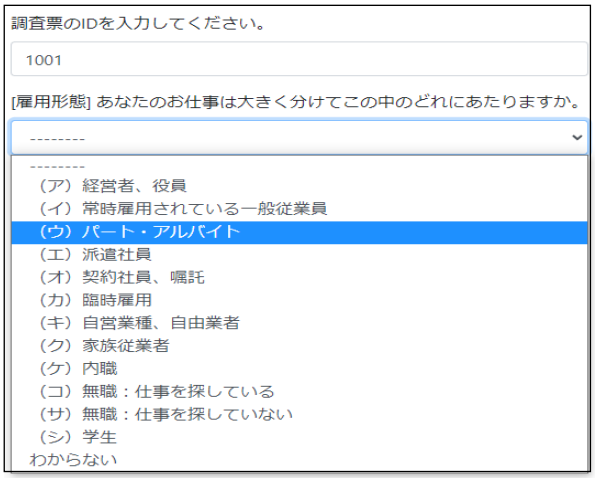
A 付録

STEP1　初期回答入力用画面の例

画面の質問文とプルダウンで表示される選択肢は，2025年 SSM 調査の調査票における「職業情報」と「学歴」を尋ねる部分を Web 用に作成し直した．入力を完了し「次に進む」ボタンを押すと，確認画面が表示され，サーバに送信される．記入漏れがある場合は入力を促すメッセージが表示される．



付録図 1　初期画面

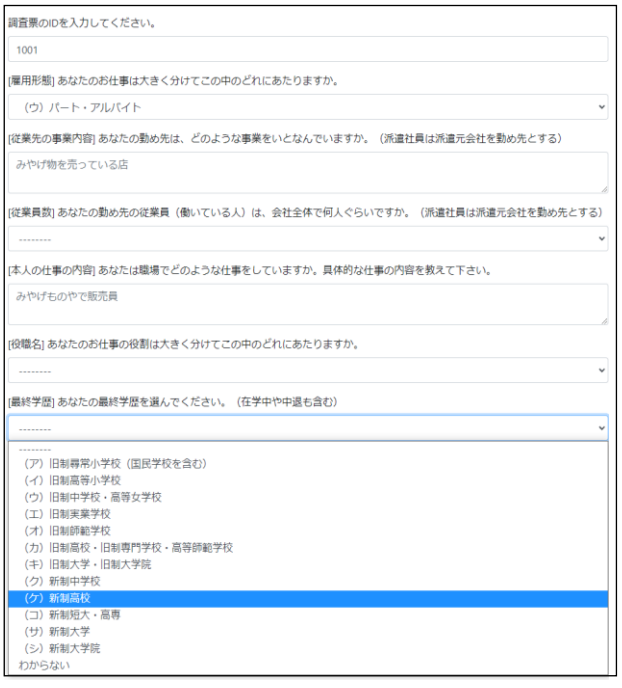


付録図 2　ID と雇用形態（「地位」）入力

付録図 3　「従業先の事業」「仕事の内容」入力

付録図 4　従業員数（「従業先の規模」）入力

付録図 5　「役職」入力



付録図 6　「学歴」入力