様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ９ 月 １４ 日現在

機関番号：３２５０２
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ０４０３９

研究課題名（和文）自然言語処理を適用した調査現場における自由回答収集支援システムの開発

研究課題名（英文）Effective Information Additional Collection System Applying Natural Language Processing at the Time of Collecting Open-Ended Survey Responses -Using Occupational Coding as Example-

研究代表者

高橋　和子（TAKAHASHI, KAZUKO）

敬愛大学・国際学部・教授

研究者番号：３０２１１３３７

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：機械学習と自然言語処理の適用により、調査員が回答者から収集した自由回答を現場でタブレットに入力してクラウドサーバに置いたシステムに送信すると、分類コード決定のための情報が回答に不足しているか否かを判定し、不足する場合は有効な情報を回答者に提示し、追加してもらうシステムを構築中である。情報不足の判定は、自動コーディング結果に付与された「確信度」が最も低いレベルの場合とし、情報提示方法を4種類提案した。
アルゴリズムはほぼ完成し、性質の異なる2つのデータセットによる実験の結果、すべての提示手法で有効性が示された。今後の課題は、コーダによる評価を行うことと、実装を急ぎ、利用環境を整えることである。

研究成果の学術的意義や社会的意義
本研究は、社会調査とくに階層移動研究において必須の作業である「職業コーディング」を担当するコーダの作業負担軽減および、コーダとコーダがその結果を参考にするコーディング自動化システムの正解率向上を目的とするシステムの開発であり、社会学分野における貢献が最も大きい。
次に、ルールベース手法や機械学習（サポートベクターマシン）、自然言語処理の適用により、回答に不足する情報を調査現場で回答者自身から追加してもらうという発想は、情報処理分野の研究者でなくてはできないもので、学際的な研究であると評価できる。
さらに、情報処理分野の応用研究として社会学分野の問題解決を行うという学術的な意義もある。

研究成果の概要（英文）：We are developing a new system that applies Support Vector Machines and Natural Language Processing. In the proposed system, a survey taker first enters answers including an open-ended response collected from a respondent into a tablet and sends it to a cloud server where the system is installed. The system then classifies the open-ended response into a valid code attached the confidence level, and presents effective words according to the prediction code to the respondent if it perceives that a response does not contain sufficient information for classification. This decision is determined when the confidence level is the lowest. After collecting information selected by the respondent, the system reclassifies a new open-ended response into a valid code.
The system has not been completed yet, but shows efficacy by some experiments.
In future work, we will completely implement the system and evaluate it by coders.

研究分野：自然言語処理

キーワード：社会調査　自由回答　CAI調査　機械学習　自然言語処理　不足情報収集システム　職業コーディング　クラウドサーバ

１．研究開始当初の背景

これまでの科研費補助金により、職業・産業データに限定されているが、コーダが調査票をみながらすべて手作業で行っていた自由回答をあらかじめ用意されたコードに変換する「アフターコーディング」作業を、人工知能分野（機械学習と自然言語処理）の研究成果を取り入れて、自動的にコーディングを行うシステム（「職業・産業コーディング自動化システム」）を構築した。さらに、平成 25～27 年度科研費補助金では、この成果を活かし、次の①～④の機能（今回の科研費研究に関連の深い項目に限定）を追加し、システムの機能を向上させた。いずれも、表形式で入力された調査データそれぞれにコードを付けていくものである。

①対象とするコーディングの種類の拡張

社会調査における国内標準の SSM 職業・産業コード（それぞれ約 200 個、20 個）だけでなく、ILO が定めた国際標準コード（職業の場合は ISCO、産業の場合は ISIC）への自動変換も可能にした（図 1 参照）。

ISCO（小分類約 400 個）：International Standard Classification of Occupations

ISIC（亜大分類約 60 個）：International Standard Industrial Classification of All Economic Activities

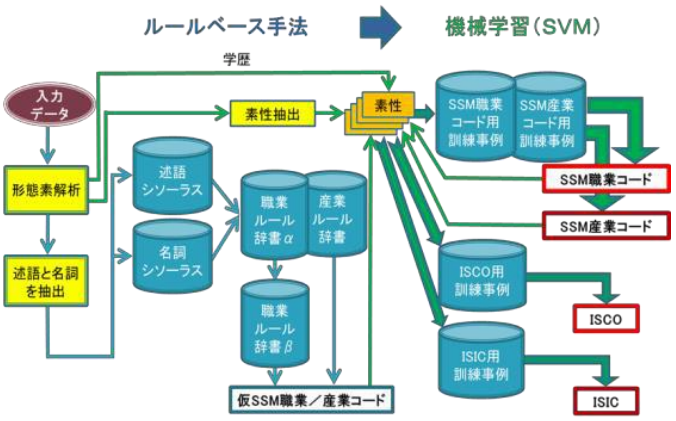

図 1　職業・産業コーディング自動化システムの構成

②システムが予測した職業・産業コードに対する確信度の付与（3 レベル）

システムが第 1 位に予測した自動コーディング結果に対し、機械学習（SVM）により出力される分離平面からのスコアにおいて、第 1 位のスコアと第 2 位のスコアの正負および、第 1 位のスコアと第 2 位のスコアの差が閾値以上か否かにより、3 段階（A「人間がチェックする必要がない（完全自動化）」、B「チェックした方がよい」、C「人間がチェックをする必要がある」）の確信度を付与する機能を追加した。

③正解率の向上

機械学習では訓練データ（正解が付いたデータ）の量が増えるほど精度が向上するため、訓練データを増強した。また、正解率向上のために、SSM 産業コードや国際標準コードを予測するアルゴリズムの開発も行った。

④Web 版システムの公開

本システムを東京大学社会科学研究所附属社会調査・データアーカイブセンターSSJDA（後の CSRDA）に置き、図 2 に示す方法により、Web による利用公開を実現した。

(1)[利用者]利用申請書をメイルにより SSJDA に送信（希望する職業・産業コードの種類を明記）

(2)[SSJDA]ユーザ ID、パスワードの発行とアップロード（ダウンロード）場所の指定

(3)[利用者]入力用データファイルをアップロード

(4)[利用者]結果ファイルをダウンロード

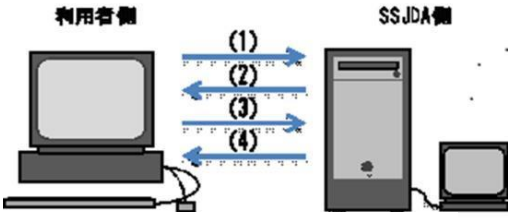

図 2　Web 版システムの利用方法

セキュリティの問題で、完全なオンラインではなく、(3)と(4)の間で、CSRDA の運用担当者（社会学研究者）がオフラインで操作する。このため、誰もが容易に操作できる画面とした

（図 3 参照）。

本システムは、前述の機能拡張を行ったことで、社会学研究者は、大規模調査であっても、各自の研究室にいながら、4 種類の職業・産業コーディング結果を確信度付きで入手できるようになった。実際、この Web 公開により 12 件の利用があった（その他に、2015 年 SSM 調査における職業コーディングでも約 6 万事例で利用された）。また、本研究の研究成果として、社会調査方法論と情報処理分野にまたがる学際的な研究であるために、社会調査における他分野との協同の発展可能性を広げたことも挙げられる。

図 3　CSRDA 担当者操作画面（初期画面）

以上に述べたように、本研究は、社会学分野や情報処理分野に貢献したと評価できる。しかし、本システムを構築した本来の目的は、作業負担の多い「コーダの支援」であり、その観点からは、なおも次の課題が存在する。これは、豊田・常松（2008）による「コーディング作業における留意点」（『職業・産業コーディングマニュアルと作業記録』（田辺・相澤編）pp.53-58）の「コーダアンケートから見る傾向」でも報告されているように、コーディングの際、「回答の内容から仕事の重心や内容が選定し辛い場合」すなわち、分類のための判断材料とする情報が回答に不足している中で何らかのコードを付けなくてはいけない状況があることである。これは、コーダを悩ませ、コーダの作業負担感を増す問題で、職業・産業コーディングに特有なものではなく、カテゴリの準備された自由回答が抱える重要な問題であると考えられる。したがって、次の課題は、「職業・産業コーディング」を手がかりにしながら、この解決方法を検討することである。

２．研究の目的

本研究の目的は、自由回答のアフターコーディングを行うコーダ支援として、まずは、１．で述べた現自動化システムの構築では解決できなかった問題、すなわち、アフターコーディング時に、分類するための情報が十分含まれていない回答に対するコーダの負担感を軽減化する必要があるために、これに対する支援方法を検討し、解決に向けた方法を提案することである。次に、この目的で構築したシステムを誰もが自由に利用できる仕組みをどのように作ればよいかについても検討することである。さらに、上記 2 つについて、アフターコーディングが必要な自由回答全般が抱える問題であることを認識し、まずは職業・産業コーディングを対象としたシステムを構築し、その後、これを一般化することも行いたい。

具体的な研究目的は、次の 5 項目である。

(1) 「職業・産業コーディング自動化システム」を整備

現自動化システムの利用を前提とする新システムの基盤を固めるため、新システムの構築を始める前に、現システムの正解率の向上と予測コードに付与される確信度の信頼性を向上させておく必要がある。具体的には、ルールベース手法で用いるシソーラスとルール辞書の整備、訓練事例の整備を行う。

(2) 調査現場で必要な情報を追加収集できる方法を提案

アフターコーディングの時点で、分類に必要な情報が回答に不足していることが明らかになっても、回答者から新たに情報を収集することが不可能である。情報不足であるか否かが調査現場で判明していれば、回答者から追加情報を得ることが可能である。

(3) 調査現場で調査員が利用するために、誰もが容易に操作可能かつその場で直ちに追加質問をして情報を収集できるシステムの利用形態を提案

(4) システム利用者による評価

システムを実際に利用するのは、調査現場で回答を収集する調査員、質問を受ける回答者、アフターコーディングを行うコーダの三者である。アルゴリズムの有効性だけでなく、実装後には、この三者によるシステムの性能についての評価を行う必要がある。

(5)システムの汎用化（カテゴリのある自由回答一般への拡張）を視野に入れた開発

本システムの対象は職業・産業情報に限定されるが、他の自由回答データに対しても適用できるシステムに容易に拡張できるような仕様にしておく。

３．研究の方法

研究の目的ごとに方法をまとめる。研究の目的(2)(3) に関わる部分は、研究代表者が考

案したアルゴリズムに基づき、プログラムの作成は業者に依頼を行う。

(1)「職業・産業コーディング自動化システム」を整備

	内　容	担当者	予定年度
①	シソーラスに追加する新語の抽出	研究代表者	平成 28 年度
②	述語／名詞シソーラスに新語を追加		
③	『職業名索引（厚生労働省）』記載情報の入力と職業ルール辞書への追加		
④	職業／産業ルール辞書の改訂		
⑤	全訓練事例の正解見直し（正解の一貫）	研究協力者（東大社会科学研究所教員 A)	平成 29 年度

(2)調査現場で必要な情報を追加収集できる方法を提案

	内　容	担当者	予定年度
①	有効情報追加収集部分の仕様検討	研究代表者	平成 28 年度
②	同上　設計	研究代表者・研究協力者 B	平成 29 年度
③	同上　プログラミング	研究代表者・研究協力者 C	平成 30・31 年度

(3) 誰もが容易に操作可能かつその場で直ちに追加質問をして情報を収集できるシステムの利用形態を提案

当初は、ノート型 PC にシステムを搭載する予定であったが、処理速度や扱いの容易さの問題から、タブレットにデータ入力してクラウドサーバに設置したシステムに送信し、結果をタブレットで受け取るオンライン利用に変更した。

	内　容	担当者	予定年度
①	システムの利用環境仕様検討	研究代表者・研究協力者 C	平成 31 年度
②	同上　利用手続き		

(4) システム利用者による評価

	内　容	担当者	予定年度
①	コーダによる評価	研究代表者・研究協力者 D・E	平成 31 年度
②	調査員･回答者による評価	研究代表者・研究協力者 F	

(5) システムの汎用化（カテゴリのある自由回答への拡張）を視野に入れた開発

	内　容	担当者	予定年度
①	システム全体の仕様検討・設計	研究代表者	平成 28～30 年度

４．研究成果

研究の目的ごとに成果をまとめる。

(1)「職業・産業コーディング自動化システム」を整備

計画通り、ルールベース手法で用いるシソーラスとルール辞書の整備、訓練事例の整備を行った。

［主な成果発表］雑誌論文①

(2)調査現場で必要な情報を追加収集できる方法を提案

図 4　調査現場のイメージ

STEP1　データ入力とサーバへの送信
調査員が回答者から得られた回答をタブレットに入力し、サーバに送信する（図 4、図 5 参照）

STEP2　自動コーディング
自動コーディングにより、回答に対する「コード」を信頼度の目安となる「確信度」（表 1 参照）を付けて予測する（図 6 参照）

STEP3　情報不足の判定
STEP2 で付与された確信度が「もっとも低いレベルの場合」（今回はE）を情報不足と判定して **STEP4** に進み、それ以外の場合は終了する

STEP4　追加情報の提示と収集
追加自動コーディングが予測したコードに応じて追加する情報の候補（単語レベル）をタブレットに提示し(提示方法は表 2 参照)、回答者により選択された情報を初期の回答に追加して STEP2 に戻る（図 7 参照）

図 5　STEP1 の入力画面

図 6　STEP1 の入力画面例（続き）と STEP2 の出力画面例

表 1　確信度レベルを決定する分類スコアの値

確信度レベル	条件
A	Score1>0 かつ Score2<=0 Score1-Score2> α
B	Score1>0 かつ Score2<=0 Score1-Score2<= α
C	Score1>0 かつ Score2>0
D	Score1<=0 Score1-Score2> β
E	Score1<=0 Score1-Score2<= β

Score1、Score2 はそれぞれ第 1 位、第 2 位に予測されたコードの分類スコアを表す。閾値 α=3、閾値 β=0.4 とした。

表 2　提示方法

方法	提示する情報
1	カテゴリ名（カテゴリ名またはこれに該当する語がある場合）
2	混同されやすいコード間で決め手となる語（これまでの経験より）
3	コードを特徴付ける複合語（訓練事例利用
4	不正解コードと正解コードを弁別する語

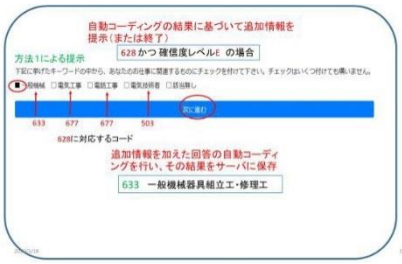

図 7　STEP4 の出力画面

本アルゴリズムを性質の異なる 2 つのデータセットにより実験した結果、4 つの方法のいずれも、語を追加しない場合より正解率が向上し、有効性を示した。今後の課題は、この手法を組みあわせた手法を提案し、有効性を示すことである。
［主な成果発表］雑誌論文②学会発表①②③

(3)調査現場で調査員が利用するには、誰もが容易に操作可能かつその場で直ちに追加質問をして情報を収集できるシステムの利用形態を提案
システムの実装は未完成であるが、「3. の（3）」で述べたように、調査員は PC ではなく、タブレットを持参する形態を想定する。［主な成果発表］学会発表④⑤

（4）システム利用者による評価
研究期間中には達成できなかったが、研究期間終了後に、まずはコーダによる評価を予定している。

（5）システムの汎用化（カテゴリのある自由回答一般への拡張）を視野に入れた開発
汎用システムの構築はできていないが、この拡張が容易なシステムである。

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 高橋和子	４．巻 Vol.7, No.1
２．論文標題 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―職業コーディングを例として―	５．発行年 2018年
３．雑誌名 日本分類学会論文誌 データ分析の理論と応用	６．最初と最後の頁 pp.21--42
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

１．著者名 高橋和子・多喜弘文・田辺俊介・李偉	４．巻 24(1)
２．論文標題 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	５．発行年 2017年
３．雑誌名 自然言語処理	６．最初と最後の頁 135-170
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 ―

〔学会発表〕　計5件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 高橋和子・奥村学
２．発表標題 機械学習の適用による調査現場での自由回答収集支援システムの構想
３．学会等名 数理社会学会第66回大会（会津大学　2018年8月30日）
４．発表年 2018年

１．発表者名 高橋　和子
２．発表標題 機械学習を適用した調査現場における追加情報収集システム―職業コーディングの場合―
３．学会等名 数理社会学会第63回大会（関西大学　2017年3月14日）
４．発表年 2017年

１．発表者名
高橋和子

２．発表標題
調査現場における自由回答収集支援システムの構想―アフターコーディング作業の効率化と精度向上に向けて―

３．学会等名
数理社会学会第64回大会（札幌学院大学 2017年9月18日）

４．発表年
2017年

１．発表者名
高橋　和子・奥村　学

２．発表標題
機械学習の適用による調査現場での自由回答収集支援システムの構築

３．学会等名
数理社会学会第68回大会（熊本県立大学　2019年8月31日）

４．発表年
2019年

１．発表者名
高橋　和子・奥村　学・鈴木　泰山・清家　大嗣

２．発表標題
機械学習の適用による調査現場での自由回答収集支援システム

３．学会等名
言語処理学会（オンライン学会　2020年3月18日）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
敬愛大学｜国際学部｜教員紹介｜国際学部国際学科｜高橋　和子教授
https://www.u-keiai.ac.jp/international/teacher/inter-study/takahashi/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考