様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １６ 日現在

機関番号：３２５０２
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３８０６４０

研究課題名（和文）社会調査の基盤を提供する自動コーディングシステムのWeb提供：その国際化と汎用化

研究課題名（英文）An Automatic Coding System for Answers to Open-ended Questions in social surveys

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）　　3,500,000円

研究成果の概要（和文）：社会調査において自由回答で収集された職業データを自然言語処理や機械学習の適用により国内／国際標準コードに自動的に変換し、その結果に人間が見直す必要性を確信度として付与するシステムを開発した。本研究ではこれをさらに発展させ、産業データの国内／国際標準コードへの自動変換、国内標準コード付きの事例に国際標準コード付与、システムメンテナンスの自動化機能を追加した。平成25年秋以降、東大社会科学研究所（CSRDA）のWebから試供提供され、利用者は入力ファイルをアップロードすれば希望するコードの結果ファイルをダウンロードできる。現在、カテゴリをもつ自由回答のコーディング自動化システムに拡張中である。

研究成果の概要（英文）：Based on the National/International standard of occupation/industry code (SSM/ISCO/ISIC), the system can assign three candidate codes to an answer of an open-ended question in a social survey. And a three-grade confidence level was assigned to the first-ranked predicted code by using classification scores. Many functions are implemented in this system, for example, an ISCO/ISIC code can be assigned to the SSM code automatically, also maintenance function contained. This system is now released to the public. Researcher can use it through the official website of CSRDA.
A new automatic coding system for social surveys is in development, which contains an open-ended question.

研究分野：自然言語処理、機械学習、社会調査法

キーワード：社会調査　自由回答　職業／産業コーディング自動化システム　SSM/ISCO/ISIC　Web公開システム　機械学習　自然言語処理　確信度

１．研究開始当初の背景

科研費補助金（平成 7 年度、平成 16～17 年度、平成 22～24 年度）により、職業データに限定されてはいるものの、これまで手作業で行っていた自由回答から基礎データへの変換を、人工知能分野の最新の研究成果を取り入れて迅速に正確に随時提供できる「職業コーディング自動化システム」を構築した。特に、ISCO（International Standard Classification of Occupations）への変換は国際比較研究の推進により、社会学の進展に大きく貢献する。本システムの Web による公開も提案したが、このサービスが開始すれば、個々の研究者の調査だけでなく大規模調査のコーディングも各自の居場所で各自の予定に合わせた実施ができる。具体的には下記 4 項目の研究を進め、課題を残していた。

(1)「職業データ」の国内／国際標準コードへの自動コーディング

システムは、自由回答で収集された「職業データ」に対して、ルールベース手法の結果も機械学習の素性として利用する手法の適用により、国内の社会調査において標準コードとされている「SSM 職業コード」（約 200 個）および ILO が定めた国際標準コードである「ISCO」（約 400 個）に自動的に変換する（図 1 参照）。

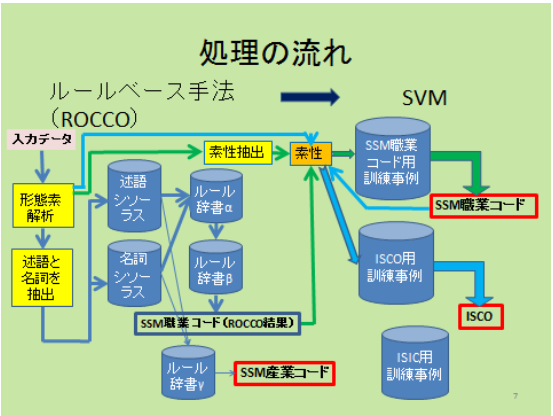


図 1　職業コーディング自動化処理の流れ

しかし、産業データは、ルールベース手法により国内標準である「SSM 産業コード」に自動変換するが、機械学習が適用されていないため、(2)で述べる確信度の付与ができない。また、国際標準産業コードである「ISIC」International Standard Industrial Classification of All Economic Activities）への自動変換を行うことができない。

［課題］産業データに対しても、機械学習の適用による国内／国際標準コードへの自動変換を行う。これにより、職業・産業データは国内／国際標準の計 4 種類のコードに自由に変換できる機能をもつシステムにする。

(2) 予測した「職業コード」に対する確信度の付与

システムは、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 位のスコアの正負、第 2 位のスコアの正負、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もつ。

確信度別の分類精度（以下、正解率とよぶ）と再現率は表 1、表 2 の通りである。ここで、正解率とは、システムが正しいコードを付けた事例を全事例で割った値、再現率とは、システムがその確信度を付けた事例を全事例で割った値である。

表 1　確信度別正解率（第 1 位のみ）

コード	A	B	C
SSM 職業コード	95%	72%	36%
ISCO	94%	68%	28%

表 2　確信度別カバー率（第 1 位のみ）

コード	A	B	C
SSM 職業コード	29%	43%	28%
ISCO	7%	67%	26%

［課題］「産業コード」に対しても確信度を付与する。

(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。SSM 職業コーディング自動化システムにおける訓練データとして、これまで利用してきた JGSS-2000、-2001、-2002、-2003 データセットに JGSS-2005 データセットを追加した（計 39,120 サンプル）。SSM コードにおける分類精度（第 3 位まで）は表 3 の通りである。ただし、SSM 産業コードには機械学習が適用されていないため、訓練事例の追加による影響は受けない。

表 3　コードの種類別正解率（第 3 位まで）

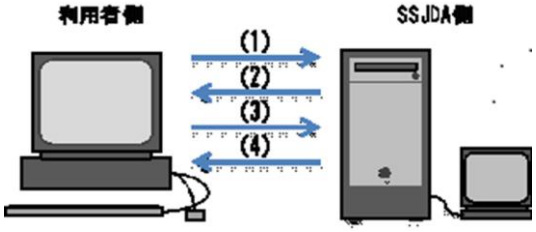
コードの種類	JGSS-2006	EASS-2006	JGSS-2010	2005 SSM
SSM 職業	79%	79%	78%	81%
SSM 産業	71%	73%	74%	70%

［課題］正解率のさらなる向上を目指す。

(4) Web 版システムの提案

Web 版システムとして、図 2 に示す方法を

提案した。システムは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、(3)と(4)の間で SSJDA の運用担当者が操作することを想定し、だれもが容易に操作できる画面とした。



(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　システムの利用方法（提案）

[課題] Web 版システムを公開できる仕組みを作り、公開を実現する。

２．研究の目的

本研究の目的は、１．で述べた「職業コーディング自動化システム」における諸課題を解決し、近年の社会学における「国際化」に対応するため、産業データについても職業データと同様の自動変換や確信度付与の機能を追加することである。また、システムの永続性の観点から、開発者以外のだれもが容易にメンテナンスを行うことができる機能の追加も行う。一方で、自動コーディングの対象を職業や産業データに限定せず、自由回答一般に拡張した汎用システムの構築も行う。具体的には、下記 6 項目を研究目的とする。

(1) 「産業データ」の国内／国際標準コードへの自動コーディング機能の追加

産業データに対しても、機械学習の適用により、「SSM 産業コード」および国際標準コード「ISIC」への自動変換機能を追加する。これにより、職業・産業データは国内／国際標準コードの計 4 種類のいずれにも自由に自動変換することができ、また機械学習を適用することで、いずれのコードにも確信度を付与することができる。

(2) 「過去の調査ですでに国内標準コードが付与されている職業・産業データ」の ISCO／ISIC への自動コーディング機能の追加

国際標準のコードへの要請は高く、国内標準コードの国際標準コードへの変換も求められている。しかし、現在は両者の間に単純な対応関係が見出しにくい状況となっているため、すでに付けられた国内標準コードを利用した国際標準コードへの自動コーディング機能を追加し、この結果にも確信度を付与する。

(3) 正解率のさらなる向上

実験の結果、ルールベース手法による正解率が高いほど機械学習の正解率も高いことがわかったため、ルールベース手法の見直しを行い、シソーラスやルール辞書を改善する。

(4) Web 版システムとしての試行提供開始

CSRDA の Web を通じた利用を可能にするために、図 2 を実現させる。

(5) システムのメンテナンス処理の自動化

今後もシステムが永続的に利用されるには、開発者以外の人間でもメンテナンスを継続していくことができることが必要である。その中でもっとも困難な作業は、調査のたびに蓄積される職業や産業の正解が付いた事例から機械学習に用いる訓練事例を生成し、既存の事例に追加することであるため、この処理を自動化する機能を開発する。

(6) 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）

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拡張する。

３．研究の方法

研究目的ごとに表にまとめる。システム再構築（(1)(2)(5)(6)）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、研究協力者がプログラムの作成を行う。

(1) 国際化（「産業データ」の国内／国際標準コードへの自動コーディング機能の追加）

表 4　研究目的(1)の研究方法

	内容	担当者	予定年度
①	SSM 職業自動コーディングのアルゴリズムを参考に、SSM 産業コードも機械学習を適用した自動コーディングおよび確信度を付与するアルゴリズム開発とプログラム作成	研究代表者 研究協力者	平成 25 年度
②	ISCO 自動コーディングのアルゴリズムを参考に、ISIC に機械学習を適用した自動コーディング機能および確信	研究代表者 研究協力者	平成 26 年度

	度を付与するアルゴリズム開発とプログラム作成		

(2) 「過去の調査ですでに国内標準コードが付与されている職業・産業データ」の ISCO／ISIC への自動コーディング機能の追加

表 5　研究目的(2)の研究方法

	内容	担当者	予定年度
①	ISCO、ISIC ともに、入力ファイルは通常のものに正解を付けた形式とし、ルールベース手法により SSM コードを予測する代わりにこの正解を機械学習の素性とする自動コーディングおよび確信度を付与するアルゴリズム開発とプログラム作成	研究代表者 研究協力者	平成 26 年度

(3) 正解率のさらなる向上

表 6　研究目的(3)の研究方法

	内容	担当者	予定年度
①	述語シソーラスと名詞シソーラスの改善	研究代表者	平成 27 年度
②	職業ルール辞書と産業ルール辞書の改善	研究代表者	平成 27 年度

(4) Web 版システムとしての試行提供開始

表 7　研究目的(4)の研究方法

	内容	担当者	予定年度
①	新システム利用手続き（必要書類など）の再検討と作成	研究代表者 共同研究者 2 名	平成 25 年度
②	新システム利用のための Web ページ作成	共同研究者 1 名	平成 25 年度

(5) システムのメンテナンス処理の自動化

表 8　研究目的(5)の研究方法

	内容	担当者	予定年度
①	正解付き事例から訓練事例を精製する処理を自動化するアルゴリズム開発とプログラム作成	研究代表者 研究協力者	平成 27 年度

(6) 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）

表 9　研究目的(6)の研究方法

	内容	担当者	予定年度
①	自由回答一般の自動コーディングへ拡張するアルゴリズム開発とプログラム作成	研究代表者 研究協力者	平成 27 年度

４．研究成果

「２．研究目的」で述べた(1)から(6)ごとに成果をまとめる。また、本システムの主な実績としては、2015 年 SSM 調査における職業コーディングでの利用（約 6 万事例）、CSRDA からの Web 公開により 12 件の利用があった。なお、数値としては示せないが、本研究は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研究成果の一つといえよう。

(1) 国際化（「産業データ」の国内／国際標準コードへの自動コーディング機能の追加）

図 3 に示すシステムを構築した。4 種類のコードにはすべて確信度が付与される。

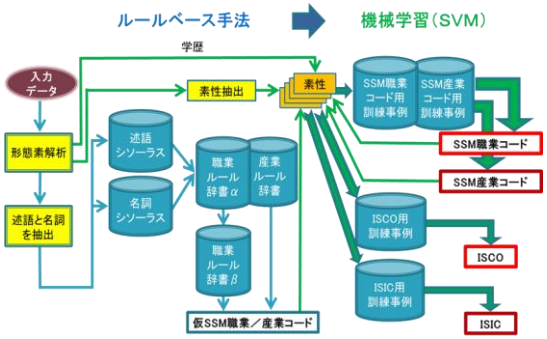


図 3　職業・産業コーディング自動化システムの構成

［主な成果発表］図書①　学会発表①⑦

(2) 「過去の調査ですでに国内標準コードが付与されている職業・産業データ」の ISCO／ISIC への自動コーディング機能の追加

本機能は特殊な場合で、図 3 には示されていない。実験の結果、この方法は通常の方法より分類精度が約 5～6%高いことが示されたため（表 10 参照）、ISCO や ISIC に変換したい事例に国内標準のコードが付与されている場合は、これを正解として入力し、システムで利用することを推奨する。

［主な成果発表］　図書①　学会発表④⑦

(3) 正解率のさらなる向上

システムを改善し、現在の正解率は表 10 の通りとなった。表中、ISCO*や ISIC*は、(2) で述べた場合の結果である。今回、SSM 産業コードで正解率が大きく向上したのは機械学習を適用した効果によると考えられる。

表 10　コードの種類別分類精度（正解率）

	コードの種類	第 3 位まで
国内標準コード	SSM 職業コード	約 80%
	SSM 産業コード	約 90%
国際標準コード	ISCO	約 70%
	ISIC	約 80%
	ISCO*	約 75%
	ISIC*	約 86%

なお、当初の予定にはなかったが、本システムを実際に利用する社会学者（共同研究者）による評価も行った。実際の分析で用いる各種大分類レベルでは特に確信度Ａが付与された事例で有効性が認められた（システムの結果は人手を介さずそのまま利用可能）。

［主な成果発表］図書① 学会発表④⑤⑦⑧

(4) Web 版システムとしての試行提供開始

システムは平成 25 年 9 月より CSRDA の Web サイトより公開された。CSRDA 担当者用の操作画面を図 4 に示す。

図 4　CSRDA 担当者操作画面（初期画面）

［主な成果発表］図書① 学会発表②④⑥⑦その他②

(5) システムのメンテナンス処理の自動化

訓練事例の追加が自動的に行える機能も追加したバージョンの操作画面を図 5 に示す。

図 5　訓練事例の追加自動処理機能を追加したシステム（初期画面）

［主な成果発表］図書① 学会発表⑥⑦

(6) 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）

職業・産業コーディング自動化システムを 3 種類の選択回答と 2 種類の自由回答からなる回答を総合的に判断して分類するタスクと捉え、システムの拡張を行っている。操作画面の設計と自動化のアルゴリズムは完成したがプログラムが未完であるため、研究期間終了後も継続して完成させる。

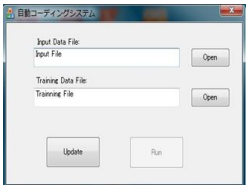


図 6　カテゴリのある自由回答一般への拡張システム（初期画面）

［主な成果発表］図書① 学会発表⑦

５．主な発表論文等

〔雑誌論文〕（計　１件）

① 多喜 弘文、専門学校の位置づけとその変化における男女差―「変容モデル」の批判的検討、『全国無作為抽出による「教育体験と社会階層の関連性」に関する実証的研究』（中村高康編）、査読なし、2016、196-212　DOI: なし

〔学会発表〕（計　８件うち国際会議　１件）

① 高橋 和子、多喜 弘文、田辺 俊介、李 偉、社会調査における職業・産業コーディング自動化システムの一般公開と運用、言語処理学会第 20 回年次大会論文集、932-935、2014 年 3 月 20 日、北海道大学 *http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/P8-15.pdf*

② 高橋 和子、多喜 弘文、田辺 俊介、李 偉、職業・産業コーディング自動化システムの一般公開に向けた課題と対応、数理社会学会第 57 回大会報告要旨集、68-71、2014 年 3 月 8 日、山形大学

③ 多喜 弘文、学歴としての専門学校に関する計量的研究―ESSM2013データを用いて―、日本教育社会学会、2014年9月30日、愛媛大学・松山大学

④ Kazuko TAKAHASHI and Hirofumi TAKI and Shunsuke TANABE and Wei LI, October 22, 2014, “An Automatic Coding System with a Three-Grade Confidence Level Corresponding to the

National/International Occupation and Industry Standard: Open to the Public on the Web", *Proceedings of the 6th International Conference on Knowledge Engineering and Ontology Development (KEOD 2014 in Roma)*, 369-375, DOI: 10.5220/0005131703690375
https://www.researchgate.net/publication/276090302_An_Automatic_Coding_System_with_a_Three-Grade_Confidence_Level_Corresponding_to_the_NationalInternational_Occupation_and_Industry_Standard_Open_to_the_Public_on_the_Web（論文の一部）

⑤高橋 和子、職業コーディング自動化システム評価（得意／苦手な分類）、数理社会学会第 59 回大会報告要旨集、76-79、2015 年 3 月 14 日、久留米大学（福岡県・久留米市）

⑥ 高橋 和子、多喜 弘文、田辺 俊介、李 偉、機械学習を適用した自由回答のコーディング支援―職業・産業コーディング自動化システムとその拡張―、情報処理学会第 78 回（平成 28 年）全国大会講演論文集（４）、495-496、2016 年 3 月 10 日、慶應大学矢上キャンパス（神奈川県・横浜市）

⑦ 高橋 和子、多喜 弘文、田辺 俊介、李 偉、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―自然言語処理と機械学習の適用―、言語処理学会第 22 回年次大会ワークショップ「言語処理の応用」、2016 年 3 月 11 日、仙台国際センター（宮城県・仙台市）
http://www.anlp.jp/proceedings/annual_meeting/2016/workshop1/pdf/ws1.pdf

⑧ 高橋 和子、多喜 弘文、田辺 俊介、職業・産業コーディング自動化システムの利用に関する評価－社会階層研究を事例に」『数理社会学会第 61 回大会報告要旨集、31-34、2016 年 3 月 17 日、上智大学（東京都・千代田区）

［図書］（計　１件）
① 高橋 和子、『職業・産業コーディング自動化システム　平成 25～27 年度科研費成果報告書』、2016 年 3 月、120 頁、DOI:なし

〔その他〕（計　３件）
① ホームページ：敬愛大学＞国際学部＞教員紹介－国際学科－高橋和子
http://www.u-keiai.ac.jp/teacher/international/inter-study/takahashi/index.html

②ホームページ：東京大学社会科学研究所附属社会調査・データアーカイブ研究センター＞社会調査＞共同調査と共同研究＞自動コーディング（職業・産業）
http://csrda.iss.u-tokyo.ac.jp/joint/autocode/
同上＞自動コーディングとは？
http://csrda.iss.u-tokyo.ac.jp/joint/autocode/about/
同上＞利用方法
http://csrda.iss.u-tokyo.ac.jp/joint/autocode/usage/
同上＞入力ファイルの形式
http://csrda.iss.u-tokyo.ac.jp/joint/autocode/format/

③ ソフトウェア開発「職業・産業コーディング自動化システム」（名称 aucs.exe）

６．研究組織
(1)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

(2)研究分担者
多喜　弘文（TAKI HIROFUMI）
東京大学・社会科学研究所・助教（H24→H25）
法政大学・社会学部・講師（H26）
研究者番号：80455774
田辺　俊介（TANABE SHUNSUKE）
早稲田大学・文学学術院・准教授
研究者番号：30451876

(4)研究協力者
李　偉（RI I）
東京工業大学大学院・理工学研究科・博士課程在学