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を適用して—

高橋 和子
敬愛大学国際学部
takak@u-keiai.ac.jp

多喜 弘文
法政大学社会学部
taki@hosei.ac.jp

田辺 俊介
早稲田大学文学学術院
tanabe.sh@waseda.jp

李 偉
東工大大学院理工学研究科
li.w.aa@m.titech.ac.jp

1 はじめに

社会学においては, 職業や産業データは性別や年齢などと同様に重要な属性であり, 正確を期する必要がある. このため, 国勢調査でも行われているように, 自由回答で収集したものを研究者自身の手で職業 · 産業分類コードに変換する場合が多い [4]. この作業は「職業 · 産業コーディング」とよばれるが, 職業小分類コードは約 200 個, 産業大分類コードでも約 20 個あり, 分類すべきクラスの数が非常に多いこと, コード化のルールも複雑なことから, 特に大規模調査の場合は多大な労力や時間を要するという深刻な問題を抱えている [7]. また, 多人数で長期間にわたる作業となるため, コーディング結果における一貫性の問題も存在する [23]. そこで, これらの問題を軽減する目的で, 職業 · 産業コーディングを自動化するシステムの開発を行ってきた.

当初は, ルールベース手法の適用により, 社会学において国内標準コードともいえる SSM 職業小分類コードと産業大分類コード [1] を対象としたシステムであった [8]. その後, システムの精度向上をめざし, 文書分類において分類性能の高さで評価されている機械学習のサポートベクターマシン（SVM）にこのルールベース手法を組み合わせた手法を開発した [11].

一方で, 社会学における国際比較研究の隆盛に対応するため, 処理の対象とするコードを, ILO により定められた国際標準職業分類 ISCO（International Standard Classification of Occupations）や国際標準産業分類 ISIC（International Standard Industrial Classification of All Economic Activities）[3] とするシステムも開発した [13]. 社会学では, ISCO は小分類コード, ISIC は亜大分類コードが用いられることが多く, その個数はそれぞれ約 400 個と約 60 個である. SSM 職業 · 産業分類は, もともとは 1968 年版 ISCO や ISIC を源とする日本標準職業 · 産業分類を社会学で使用しやすいように改変されたものであったが, その後の ISCO と ISIC における大幅な改訂の結果, 両者の対応関係は複雑化した [21]. これが ISCO や ISIC のために新規に自動化システムを開発した理由である. なお, ISCO や ISIC のためにはルールベース手法は構築していない.

自動化システムは, いずれも入力ファイル（CSV 形式）にある職業や産業情報から予測したコードを CSV 形式で提示するため, コーダはこれを参考にしながらコーディングを行える. 特に初心者のコーダに対する有効性が評価され, 我が国初の二次分析のための大規模調査 JGSS（Japanese General Social Surveys；日本版総合的社会調査）[1] において, 初回の 2000 年以降, 毎回利用されてきた [9] [10] [12] [15]. また, 10 年ごとに実施される SSM（Social Stratification and socila Mobility）調査（社会階層と社会移動全国調査）においても, 2005 年調査に引き続き [13], 2015 年調査 [2]でも利用されている. SSM 調査は, 社会学の中でも職業や産業データがとりわけ重要な役割を果たす階層移動研究の調査で, 大規模である上に, 本人の初職から現職にいたるまでの職業や産業の履歴に加え, 配偶者, 父親, 母親についても収集されるため, 作業量の問題は重大である. 例えば 2015SSM 調査の場合, コーディングを行う事例は約 60,000 にのぼっている

自動化システムを利用することにより, コーダの作業内容は楽になったが, すべてのデータに対して作業をすることについては変わりがない. そこで, コーダの作業の絶対量についても軽減できるように, 自動コーディング後にコーダの作業が必要かどうかを示す目安として, 3 段階（A:不要, B：できれば必要. C:必要）の確信度を付与する機能の追加を行った [16].

このように, 自動化システムは機能を充実させながら, JGSS や SSM 調査のような大規模調査で利用されてきた. しかし, これ以外にも, 研究者グループや個人からの依頼を受けて開発者自身が処理を行ってきたケースもあり, 利用者が増えるにつれ, 負担となっていた. そこで, Web を通じてだれもが自由に自動化システムを利用できる仕組みを検討した. その際, 利用者の多くが文系の研究者であると予想されることや, システムの稼

[1] http://jgss.daishodai.ac.jp/surveys/sur_top.html
[2] http://www.l.u-tokyo.ac.jp/2015SSM-PJ/index.html

働環境がやや複雑であることから, 利用者自身がシステムをダウンロードして用いるのではなく, 入力データのファイルをアップロードしたものをシステム運用担当者が処理する方法を想定した. このとき, システム運用業務もだれもが担当できるよう, ユーザーインタフェイスを重視した [17]. またこれを機に, これまで開発してきた種々の自動化システムの整理統合を行うことにした.

以上の方針のもとに改造された自動化システムは, 現在, 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（SSJDA）[3]に置かれ, Web を通じた利用サービスが試行提供中である [18]. ここでは, Web を通じて SSJDA に利用申請した書類が受理されれば, 所定の形式の入力データファイルを指定された場所にアップでき, 希望する職業や産業のコーディング結果をダウンロードできる仕組みとなっている. これにより, 一般の研究者グループや個人も利用しやすくなった.

自動化システムは現在も改良を続けているが, 本稿では, 現時点において公開されているシステムについて, その運用方法や利用方法を含めて報告する. その際, 社会学研究者の立場からのシステム評価についても簡単に報告する. 以下, 次節では, 関連研究として, 海外の職業・産業コーディングの自動化システムや公開方法について述べる. 3 節で本システムについて説明し, 4 節で評価を行う. 5 節で本システムの運用・利用方法について説明し, 最後に, まとめと今後の課題について述べる.

2 関連研究

職業・産業コーディングは, 海外においても負担の大きい作業であるとの認識があり, 本システムとは適用場所が異なるが, 自動化システムが開発されている.

韓国では, 大韓民国統計庁の Web-based AIOCS（A Web-based Automated System for Industry and Occupation Coding） [5] があり, ISCO や ISIC に由来する職業コード（442 個）や産業コード（450 個）に変換する. 利用方法は, Web サイト上に, 会社名（自由回答）, ビジネスカテゴリ, 部門, 役職, 仕事の内容（自由回答）を入力すると, 同一画面に結果が表示される. 自動化の手法は, 処理時間の問題から SVM は用いず, ルールベース手法, 最大エントロピー法（MEM）, 情報検索技術（IRT）の 3 種類を単独またはルールベース手法と組み合わせた計 6 種類である. 正解率が最も高いのは, すべての手法を組み合わせた手法（ルールベース手法, MEM, IRT をこの順に実行）で 76% である.

米国では, CDC（Centers for Disease Control and Prevention）の Web サイト上に SOIC（Standardized Occupation & Industry Coding）システムが公開され [4], 利用者自身によりソフトウェアをダウンロードして処理を行う. 自動化の手法は, ルールベース手法によるマッチングが主で, 正解率は, 職業コードでは 75%, 産業コードでは 76% で, 両方正解では 63% である.

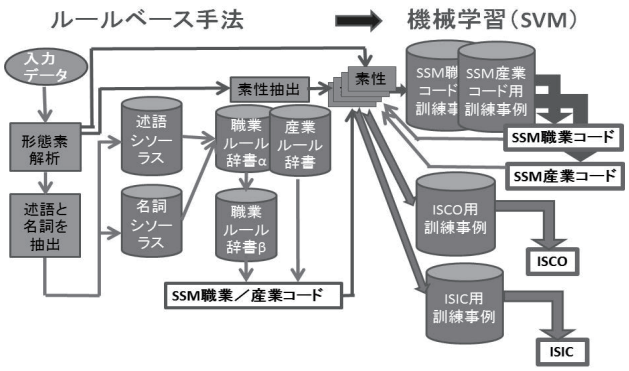

図 1: システムの構成図

CDC では, 2013 年に, SOIC の後継として新たに NIOCCS（The NIOSH Industry & Occupation Computerized Coding System）を公開し [5], 2000 年以前のセンサス・コードには SOIC. 2000 年以降のコードには NIOCCS を対応させている. NIOCCS はシステムを公開しておらず, 一問一答方式の他に, 入力情報は異なるものの, 本システムと同様にファイルによるデータの受け渡しも行える. また, 自動コーディングの結果に精度に関する確信度を 3 段階（High, Medium, Low）で付与する点も本システムと類似している.

3 職業・産業コーディング自動化システム

ここでは, 現在, SSJDA により利用が公開されている自動化システムについて述べる.

最初に, 本システムの構成を図 1 に示す. 本システムが処理するコードは表 1 に示す 4 種類で, いずれも現在の社会学で必要性が高いものである.

図 2 は, SSJDA において, 本システムの稼働時に表示される操作用の画面である. システムの運用担当者は, この画面にしたがって利用者からアップされた入力データファイルを指定し, 希望のあったコードのチェックボックスをクリックすればよい.

[3] http://csrda.iss.u-tokyo.ac.jp/joint/autocode/

[4] http://www.cdc.gov/niosh/soic/SOIC.About.html

[5] http://wwwn.cdc.gov/niosh-nioccs/

表 1: 対象とするコードの種類と個数

コードの種類	コード数	備考
SSM 職業（小分類）	約 200	1995 年版 [1]
SSM 産業（大分類）	約 20	1995 年版 [1]
ISCO（小分類）	約 400	階層構造 4 桁利用
ISIC（亜大分類）	約 60	同上 2 桁まで利用

図 2: SSJDA 運用担当者用操作画面

3.1 入力データファイルと結果ファイルの形式

入力データは, 各列が「ID」「学歴」「従業上の地位 · 役職」「仕事の内容」「従業先の事業内容」「従業先の規模」の順である CSV 形式のファイルである. このうち, 自由回答は「仕事の内容」と「従業先の事業内容」で, 他は選択回答である. 利用者は, この形式のファイルを用意すれば, 表 1 に示すコードのうち希望するコードを最大 4 種類まで自由に選択できる.

この他, 過去の調査等ですでに SSM コードが付与されているものに対して新たに ISCO や ISIC を付与することもできる. このためには, 入力データファイルとして, 前述のものの最右列（従業先の規模）の右列に「付与ずみの SSM コード」を入力したものを用意し, 希望するコードを ISCO や ISIC とすればよい. システムは自動的に決定ずみの SSM コードを利用する.

本システムでは, 結果ファイルとして, コードの種類ごとに, 第 3 位までに予測したコードを提示した CSV 形式のファイル（各列が「ID」「確信度」「第 1 位に予測されたコード」「第 2 位に予測されたコード」「第 3 位に予測されたコード」の順）を出力する.

3.2 自動化の手法

表 2 に, 対象とするコードごとの自動化の手法を示す. ISCO や ISIC において, SVM の素性として用いられる SSM 職業コードや産業コードは, ルールベース手法と SVM の組み合わせ手法により決定されるものである. 以下, 説明する（図 1 を参照のこと）.

表 2: 自動化の手法

コード	自動化の手法（SVM で用いる素性）
SSM 職業コード	ルールベース手法と SVM の組み合わせ（基本素性, ルールベース手法の結果）
SSM 産業コード	ルールベース手法と SVM の組み合わせ（基本素性, ルールベース手法の結果）
ISCO	SVM（基本素性, 学歴, 第 1 位に予測された SSM 職業コード*）
ISIC	SVM（基本素性, 第 1 位に予測された SSM 産業コード*）

＊：過去の調査等ですでに付与されたコードがある場合は, 予測コードではなく付与されたコードを用いる

3.3 ルールベース手法

格フレームの概念による職業 · 産業データの理解

職業 · 産業コーディングを自動化するには, 自由回答の記述内容をすべて解析する必要はなく, 分類に必要な部分のみでよいと考えられる. また, 職業データは個人の仕事内容, 産業データは従業先の事業内容を表現するという違いはあるが, いずれも基本的には動作の違いにより分類コードが決まると考え, 「格フレーム」の概念に基づいた情報の抽出を行う.

形態素解析 [6] を行った回答に対して, まず, 動作を表すものとして, 自由回答中の述語または述語相当語（以下, 述語とよぶ）を抽出する. 次に, その述語が取る格のうち, 分類に必要な格の語（以下, 名詞とよぶ）を抽出する. 例えば, 回答から「製造」や「教える」などの述語が抽出されたとき, 述語が「製造」であれば目的格「何を」, 「教える」であれば場所格「どこで」を示す語を抽出する. これより, （述語, 表層格, 名詞）の三つ組みを生成し（実際には, 述語そのものではなく後述する述語コードを用いる）, あらかじめこのような三つ組みと分類コードのペアにより構築しておいたルールベースの中からマッチするものを探し出し, 該当する分類コードを付ける. 例えば, 「アルバイトでケーキを製造している」であれば, （製造, ヲ, ケーキ）を抽出して職業コード「６４４」（パン · 菓子 · めん類 · 豆腐製造工）, 「大学で哲学を教えている」であれば, （教える, デ, 大学）を抽出して職業コード「５２４」（大学教員）が付けられる.

実際には, 職業は自由回答に記述された内容だけでなく, 選択回答である「従業先の事業規模」や「地位 · 役職」などの情報も用いて総合的に判断される [2]. したがって, この段階で付けられたコードは仮のコードであ

り, この後に必要に応じてチェック用のルールが適用され, ルールベース手法としての最終コードが決定される. 例えば, 前述の例（「６４４」や「５２４」）はそのまま最終コードとなるが, 自由回答の内容から「５５０」（会社 · 団体の管理職員）が付けられていても, 「従業先の事業規模」が 20 人以下であったり, 「地位 · 役職」が係長や主任などの場合は「５５４」（総務 · 企画事務員）に変更されるケースはしばしばある.

なお, ルールベース手法により決定されたコードは, 後述するように, ルールベース手法と SVM の組合せ手法においては, SVM の素性として扱われる.

シソーラスによる語の拡張

自由回答に出現するあらゆる語から生成される（述語, 表層格, 名詞）の三つ組みに対してルールを生成するのは不可能である. したがって, ルールは代表的な語により生成しておき, 述語, 名詞はそれぞれシソーラスを生成して語の拡張を行う.

述語シソーラスは, 語や品詞が異なっていても, 職業や産業コードに分類する観点からは同一とみなすことのできる語同士には同一の述語コードを付けてグループ化したものである. 例えば, 「製造」（サ変名詞）, 「製作」（サ変名詞）, 「作る」（動詞）はすべて同一の述語コード「386 1」が付けられる.

名詞シソーラスも同様で, ルールで用いられる抽象度の高い語を見出し語としたより具体的な語からなるグループである. 例えば, 見出し語が「電気機械器具」の場合, そこには「テレビ」「コンピュータ」などが含まれる. このようにして, 「コンピュータの製造」と「テレビを作る」はいずれも, （386 1, ヲ, 電気機械器具）と「６３４」（電気機械器具組立工）をペアとするルールにより, 同じ職業コード「６３４」が付けられる.

3.4 ルールベース手法と SVM の組み合せ手法

ルールベース手法と SVM の組み合わせ方にはさまざまな方法が考えられる. SSM 職業コードによる実験の結果, ルールベース手法による結果を SVM の素性として利用する方法がもっとも精度が高かったため [11], 本システムでもこの方法を適用する.

ルールベース手法をもたない ISCO や ISIC では, これと源を同じくする SSM コードをルールベース手法による結果として扱う. このとき, ISCO における実験の結果より, 予測された SSM 職業コードは第 3 位まで利用するより, 第 1 位のみの方が有効であったため [13], 第 1 位のみを利用する. ISIC においても同様にする.

表 2 における SVM の素性のうち, 基本素性とは, 「従業先事業の種類」「仕事の内容」「地位 · 役職」である. また, ISCO においては「学歴」も用いる理由は, ISCO ではコードの決定時に, 職業の遂行に必要なスキルレベル（＝教育 · 職業資格）が用いられるが, 我が国ではスキルレベルそのものは収集されないため, これが国際標準教育分類（ISCED）と対応していることや学歴を判断基準とする点 [22] に注目し, 学歴で代用可能であると判断したためである.

職業 · 産業コーディングは多値分類のタスクであるため, 2 値分類器である SVM を one-versus-rest 法により多値分類器に拡張した.

4 種類のコードをすべて提示する場合は, STEP １～STEP ６の順に処理が行われる. ISCO のみの場合でも, STEP １～STEP ３, ISIC のみの場合でも, STEP １, STEP ２, STEP ５が実行される.

STEP １　職業 · 産業情報に対する形態素解析
STEP ２　ルールベース手法の適用により, 仮 SSM 職業コードと仮 SSM 産業コードを決定
STEP ３　基本素性に, STEP ２により決定された仮 SSM 職業コードを追加して SVM を適用し, SSM 職業コードを決定
STEP ４　基本素性に, 学歴と STEP ３により決定された SSM 職業コード（第 1 位のみ）を追加して SVM を適用し, ISCO を決定
STEP ５　基本素性に, STEP ２により決定された仮 SSM 産業コードを追加して SVM を適用し, SSM 産業コードを決定
STEP ６　基本素性に, STEP ５により決定された SSM 産業コード（第 1 位のみ）を追加して SVM を適用し, ISIC を決定

3.5 確信度の付与

SVM は, 予測したコードとともにスコア（分離平面からの距離）も出力するため, これを利用して予測されたコードのクラス所属確率を計算する事が可能である [14]. 本システムでは厳密な確率の値までは必要としないため, [14] による手法を特徴づける「複数のスコアの利用」を取り入れた簡便な方法を用いて確信度を決定する

本システムにおける確信度は, 「A：人手によるコーディングは不要, B：できれば人手によるコーディングを行う方がよい, C：人手によるコーディングが必要」の 3 段階で, その決定条件は次の通りである. ただし, 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ともなって出力されるスコアを示

表 3: コードの種類別訓練事例と評価事例

コードの種類	訓練事例（事例数）	評価事例（事例数）
SSM 職業コード SSM 産業コード	JGSS-2000〜JGSS-2005（39,120 事例）	JGSS-2006（2,203 事例） 2005SSM 調査（16,083 事例）
ISCO ISIC	2005SSM 調査（16,083 事例）	JGSS-2006（2,203 事例）

表 4: 正解率（第 3 位に予測されたコードまで含む）

コードの種類	JGSS-06	2005SSM
SSM 職業コード	0.788	0.806
SSM 産業コード	0.908	0.916
ISCO	0.705	-
ISIC	0.801	-
ISCO*	0.748	-
ISIC*	0.862	-

す. また, α は閾値であり, 2005SSM 調査のデータセットによる実験の結果, $\alpha = 3$ とした.

A：$rank1 > 0$ かつ $rank2 <= 0$, $rank1 - rank2 > \alpha$

B:$rank1 > 0$ かつ $rank2 <= 0$, $rank1 - rank2 <= \alpha$

C：A, B 以外の場合

4 システムの評価

本稿では, 事例に対して最終的に人手で付与されたコードを「正解」として扱う.

システムの評価実験では, まず, コードの種類ごとの正解率と確信度付与の有効性, 処理時間について報告する（実験１）. ここで, 正解率は , 正解した評価事例数を評価事例数で割った値である [20]. 次に, 視点を変え, 本システムの利用が多い社会階層分野の研究者による評価として, 現在までに得られた結果を簡単に報告する（実験２）.

4.1 実験１

実験１で用いた訓練事例と評価事例を表 3 に示す. 訓練事例は公開版のものである. 最近用いられるようになった国際標準コードは 2005 年以前の JGSS データセットには付与されていないため, 社会学において伝統的に用いられてきた国内標準コードとは訓練事例が異なっている. 評価事例は現実の場面を想定し, 訓練事例より新しい事例を用いた.

正解率

コードの種類別の正解率（第 3 位に予測されたコードまで含む）を表 4 に示す. 表中, ISCO*と ISIC*は, 過去の調査等ですでに付与された正解 SSM 職業コード, 正解 SSM 産業コードをそれぞれ素性として用いた場合である（以下, 同様である）.

SSM コードは, 評価事例として 2 つのデータセットを用いたが, いずれも職業コードは約 80%, 産業コードは約 90% と安定している. ISCO や ISIC では, 職業も産業もこの値より約 10% ずつ低く, 正解 SSM コードを利用した場合でも約 5% ずつ低い. これは, ISCO や ISIC は SSM コードよりも分類クラスの数が多いことや, 訓練事例のサイズが小さいためであると考えられる.

追加実験として, ISCO の正解率向上を目的に, コード体系が階層構造であることを利用して, まず大分類（10 個）を学習させた後に, 大分類ごとに小分類を学習する方法も実験した. 第 1 位に予測されたコードについて, 階層構造を利用した方法の有効性を調査した結果, 効果あり 5 個, 効果なし 3 個, 変化なし 1 個であった（大分類が「0（Armed forces）」の場合は小分類が存在しないため除いた）. 次に, 大分類ごとに, この方法と本システムにおける手法（直接, 小分類を学習する）のうち正解率の高い方を選択して全体の正解率を算出したが, 本システムにおける手法より 0.5% しか向上せず, 両者を組み合わせた方法の有効性も認められなかった.

ISCO については, 今後, コーディングが普及するにつれ正解付きの事例が増えることが期待され, これを訓練事例に追加することでサイズが拡大するため, 正解率の向上が見込める（ISIC も同様である）. ただし, これを実現するには, だれもが容易に訓練事例を追加できる機能をシステムに追加することが必要である.

確信度の有効性

確信度は, 第 1 位に予測したコードに対して付与される. 表 5 に, コードの種類ごとの確信度別正解率とカバー率を示す. カバー率は, 確信度が付与された評価事例数を全評価事例数で割った値とする. SSM コードにおける値は, 2 つのデータセットの平均値である.

関連研究で述べた NIOCCS では, High, Medium, Low をそれぞれ 90%, 70%, 30% としている. 本システムにおいてもほぼ同様の結果であったが, 本システムの利用者は研究者を想定するため, 確信度 A の正解率

表 5: 確信度別の正解率（カッコ内はカバー率）

コード	A	B	C
SSM 職業	0.954(0.29)	0.716(0.48)	0.355(0.23)
SSM 産業	0.975(0.32)	0867(0.54)	0.437(0.14)
ISCO	0.963(0.05)	0.701(0.67)	0.276(0.28)
IISC	0.941(0.01)	0.919(0.56)	0.574(0.43)
ISCO*	0.947(0.05)	0.759(0.65)	0.300(0.30)
IISC*	1.000(0.01)	0.971(0.55)	0.671(0.44)

は特に高い必要がある. 表 5 では, 確信度Aが付与された事例の正解率は, コードの種類に関係なく 94% を上回っており, 有効性が認められる. ただし, カバー率が高いほどコーダの作業量軽減に貢献できるが, ISCO や ISIC では 10% 未満であり, 非常に低かった.

処理時間

処理時間は, PC の性能 [6]や訓練事例, 評価事例のサイズにより異なるが, 評価事例が JGSS-06 データセットの場合, STEP 1 から STEP 6 にそれぞれ 0 分, 7 分, 34 分, 7 分, 13 分, 2 分（計 63 分）を要した. 1 事例当たり, 約 1.7 秒かかる計算となる.

本システムで 1 度に処理できる事例数は最大 5000 であり, これより大きなサイズのデータセットの場合は数回に分けて処理する必要がある. この場合, 1 回の処理時間は約 2 時間半弱である.

4.2 実験２

現在, 社会階層分野の研究に対して, システムの結果をそのまま適用した場合の有効性と問題点に関する検討を行っている [19]. 対象とするコードは, 社会学でもっともよく利用される SSM 職業コードである. ここでは, 多変量解析などのより深い分析における評価ではないが, 基本的なものとして, 確信度の有効性や正解率についての検討結果を報告する.

訓練事例は, 最新版のシステム（未公開）におけるものを用いた. これは, 実験１で用いたデータセットから訓練事例の見直しのため一部を除いたものに, JGSS-2006, JGSS-2008, JGSS-2010 データセットと 2005SSM 調査データセットを加えた計 49,795 事例である. 新規に追加した事例には, 当初はなかった７００番台や８００番台のコードも 10 個含まれている. これらは既存のコードから分化させたものや, これまでは情報不足のために「９９９」（不明, 無回答）としていたものの中で得られた情報を最大限活かそうとしたものである. 例えば, 前者は, 「５９９」（会計事務員）から「７０１」（レジ · キャッシャー）, 「６７９」（大工, 左官, とび職）から「７０２」（大工）, 「６０７」（自動車運転者）から「７０６」（宅急便の配達）, 「５７８」（女中, 家政婦, 家事サービス職業従事者）から「８０１」（介護員, ヘルパー）, 「５９２」（その他のサービス職業従事者）から「８０２」（その他の医療 · 福祉サービス職業従事者）などを分化させたもので, 後者は, 「７０４」（製品製造作業者）や「７０３」（教員）などである.

[6]実験には, Intel Core i5 2500K Quad-Core Processor 3.3GHz を使用した.

表 6: 確信度別の正解率とカバー率

	A	B	C	全体
正解率	0.978	0.764	0.402	0.669
カバー率	0.14	0.52	0.34	1.00

評価事例は, 東京大学社会科学研究所が実施する「働き方とライフスタイルの変化に関する全国調査」（若年 · 壮年パネル調査；JLPS）[7]の第 1 波のうち, 本システムを利用するための項目を満たす 3,619 事例を用いた.

確信度の有効性

確信度Aが付与されたコードの正解率は, 表 6 に示すように 97.8% に達しており, ここでも確信度の有効性が確認された. これより, 「確信度Aが付与された場合は人手によるコーディングは不要」と主張することに一定の説得力があるといえる.

大分類に合併後の正解率

階層研究においては, 実際の分析に SSM 職業小分類コードをそのまま用いることはそれほど多くなく, 社会階層の観点から類似した性質をもつ職業をまとめて扱う場合がほとんどである. そこで, 次には, 約 200 の小分類コードを 16 の大分類コードに変換し, それをさらに, 階層研究において実際に用いる単位に近い 8 つのカテゴリ（以下では, このカテゴリを大分類とよぶ）に合併したものについても正解率を調査した. ここでの正解率は, 第 1 位に予測されたコードにおけるものである.

両者の正解率を比較した結果を表 7 に示す. 大分類に合併すると, 正解率の平均は 66.9% から 79.9% に上昇し, 特に「生産現場 · 技能職」では上昇幅が約 24% と大きい. 逆に, 「管理職」では小分類レベルで低い値は

[7]http://csrda.iss.u-tokyo.ac.jp/panel/JLPSYM/

表 7: 分類レベルによる正解率の比較

大分類の種類	小分類での正解率	大分類に合併後の正解率	確信度 A のカバー率
専門 · 技術職	0.778	0.855	0.24
管理職	0.571	0.571	0.0
事務職	0.607	0.754	0.11
販売職	0.688	0.777	0.11
サービス職	0.759	0.836	0.12
生産現場 · 技能職	0.571	0.811	0.08
運輸 · 保安職	0.779	0.814	0.34
農林	0.813	0.813	0.28

表 8: 大分類に合併後の確信度別の正解率

A	B	C	全体
0.986	0.869	0.618	0.799

大分類レベルでもそのままであり, また確信度 A が付与された事例もない. これより, 管理職が付与されたコードに対しては, コーダは注意する必要がある. ただし, 大分類ごとの出現度数は, 「事務職」(933)「生産現場 · 技能職」(802)「専門 · 技術職」(788)「販売職」(503)「サービス職」(378)「運輸 · 保安職」(140)「農林」(32)「管理職」(28) の順で, 管理職の出現率は 0.8% でしかなく, 全体に及ぼす影響はさほど大きくないともいえる.

「管理職」ほど低くはないが, 「事務職」「販売職」も小分類での正解率が低く, 大分類合併後も 80% に達していない. これら 3 つに共通する特徴を考えてみると, 職務が明確に限定されていない職業を多く含むことが挙げられる. この点は, 正解率が小分類レベルでそもそも高く, 資格との対応や必要な技能および職務が明確な「専門 · 技術職」と対照的である.

なお, 大分類に合併後の確信度を表 8 に示す. 表 6 と比較すると, 大分類レベルではどの確信度においても正解率が上昇しているが, 特に確信度 B では 10%, 確信度 C では 20% 以上の上昇幅となっている.

5 システムの利用方法

本システムの利用者は, 3 節で述べた所定の形式の入力ファイルを準備し, (1) 〜 (6) の手続きにしたがって自動コーディングの結果を得る.

(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 I D, パスワードの発行およびアップロード (ダウンロード) 場所の通知
(3) [利用者] 入力データファイル (CSV 形式) を指定場所にアップロード
(4) [SSJDA] 図 2 に示した操作画面において, 入力データファイルの指定とコードを指定
(5) [SSJDA] 結果ファイルを指定場所に置く
(6) [利用者] 結果ファイル (CSV 形式) を指定場所からダウンロード

セキュリティの点から, システム運用担当者は利用者からの入力データファイルを e-mail 等では受けとらず, オンラインストレージ構築パッケージ (Proself) [8]を介する仕組みとしている.

6 おわりに

本稿では. 社会学で活用されている職業 · 産業コーディング自動化システムについて, 現在, SSJDA により試行提供されているシステムを中心に, 運用 · 利用方法を含めて述べた. 職業コードの正解率は, 国内, 国際標準とも 80% に達しておらず, 正解率の向上が今後の課題である. しかし, いずれのコードにおいても, 確信度 A が付与された場合の正解率は 95% 以上で, 社会学者が実際に利用する大分類での評価では 99% となり, 確信度を付与することの有効性は確認できた.

職業や産業が変化することにより, 職業 · 産業コードも改変される. これは, 個々のコードレベルだけでなく, コード体系においてもいえることで, SSM 職業コードでは, ISCO に倣った 4 桁のコード体系が提案され, SSM 産業コードも ISIC との関係を重視し, 新たに中分類への移行が検討されている. ISCO や ISIC も, 将来は現在の 1988 年版から 2008 年版に代わるものと予想される.

このような状況において, 今後, 新規のコードが現行のものと単純な対応関係にある場合は問題ないが, そうでない場合には, 新規に生成した訓練事例をもつシステムが必要となる. また, 現時点でも, 正解率向上のためには, 信頼できる正解付き事例を訓練事例として追加していく効果は大きい. そこで, 訓練事例に関するこれらの作業をだれもが容易に行えるように, 現在, 自動化に取り組んでいるところである.

謝辞 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た研究プロジェクトである. 2005 年 SSM 調査データの利用に関して, 2015 年

[8] https://www.proself.jp/

SSM 調査研究会の許可を得た. 東大社研パネル調査プロジェクトにおける職業 · 産業コーディングの精度向上を目的として, 職業 · 産業の自由記述データの提供を受けた. 本研究は JSPS 25380640 の助成を受けた.

参考文献

[1] 1995 年 SSM 調査研究会. 2006. SSM 産業分類 · 産業分類（95 年版）.

[2] 1995 年 SSM 調査研究会. 2006. 1995 年 SSM 調査コード · ブック.

[3] Bureau of Statistics; International Labour Office. 2001. Coding Occupation and Industry. Bureau of Statistics; International Labour Office.

[4] 原純輔. 1984. 社会調査演習. 東京大学出版会.

[5] Y. Jung, J. Yoo, S-H. Myaeng and D-C. Han. 2008. A Web-based Automated System for Industry and Occupation Coding. In *Proceedings of the Ninth International Conference on Web Information Systems Engineering (WISE-08)*, LNCS, pp.443–457.

[6] 黒橋禎夫, 長尾真. 1998. 日本語形態素解析システム JUMAN version 3.61. 京都大学大学院情報学研究科.

[7] 盛山和夫. 2004. 社会調査法入門. 有斐閣.

[8] 高橋和子. 2000. 自由回答のコーディング支援について ー格フレームによる SSM 職業コーディング自動化システムー. 理論と方法 Vol.15 No.1, pp. 149–164.

[9] 高橋和子. 2002. JGSS-2000 における職業 · 産業コーディング自動化システムの適用. 日本版 General Social Surveys 研究論文集　JGSS-2000 で見た日本人の意識と行動 ［東京大学社会科学研究所資料第 20 集］, pp. 171–184.

[10] 高橋和子. 2003. JGSS-2001 における職業 · 産業コーディング自動化システムの適用. 日本版 General Social Surveys 研究論文集［2］　JGSS で見た日本人の意識と行動 ［東京大学社会科学研究所資料第 21 集］, pp. 179–192.

[11] 高橋和子, 高村大也, 奥村学. 2005. 機械学習とルールベース手法の組み合わせによる自動職業コーディング. 自然言語処理 Vol.12 No.2, pp. 3–24.

[12] 高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学. 2005. 職業コーディング支援システム（NANACO）の開発と JGSS-2003 における適用. 日本版 General Social Surveys 研究論文集［4］　JGSS で見た日本人の意識と行動, pp. 225–242.

[13] 高橋和子. 2008. 機械学習による ISCO 自動コーディング. 2005 年 SSM 調査シリーズ１２　社会調査における測定と分析をめぐる諸問題, pp.47–68.

[14] K. Takahashi, H. Takamura, and M. Okumura. 2008. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst* 19(2), pp.185–210. Springer London.

[15] 高橋和子. 2011. ISCO 自動コーディングシステムの分類精度向上に向けて—SSM および JGSS データセットによる実験の結果—. JGSS Research Series No.8:日本版総合的社会調査共同研究拠点研究論文集 [11], pp. 193-205.

[16] 高橋和子, 田辺俊介, 吉田崇, 魏大比, 李偉. 2013. 確信度付き職業 · 産業コーディング自動化システムの開発と公開. 数理社会学会第 55 回年次大会報告要旨, pp. 38–41.

[17] 高橋和子, 田辺俊介, 吉田崇, 魏大比, 李偉. 2013. Web 版職業 · 産業コーディング自動化システムの開発. 言語処理学会第 19 回年次大会論文集, pp. 769–772.

[18] K. Takahashi, H. Taki, S. Tanabe, and W. Li. 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. In *Proceedings of the 6th International Conference on Knowledge Engineering and Ontology Development (KEOD 2014)*, pp.369–375.

[19] 高橋和子, 多喜弘文, 田辺俊介. 2016. 職業コーディング自動化システム利用に関する評価ー社会階層研究を事例にー. 数理社会学会第 61 回大会報告要旨集（予定）.

[20] 高村大也. 2010. 自然言語処理シリーズ１　言語処理のための機械学習入門. コロナ社.

[21] 田辺俊介. 2006. ISCO と SSM 職業分類の相違点の検討ー国際比較調査における職業データに関する研究ノートー. 社会学論考 Vol.27, pp. 53–78.

[22] 田辺俊介. 2008. SSM 職業分類と ISCO-88 の比較分析. 2005 年 SSM 調査シリーズ１　2005 年 SSM 日本調査の基礎分析ー構造 · 趨勢 · 方法ー, pp.31–45.

[23] 轟亮, 杉野勇. 2013. 入門社会調査法. 法律文化社.