Web版職業・産業コーディング自動化システムの開発

高橋 和子(敬愛大学)　田辺 俊介(東京大学)　吉田 崇(静岡大学)
魏 大比(名校教育グループ)　李 偉(東京工業大学)

研究の背景

社会調査において**自由記述**により**職業・産業**情報を得た場合はコーディング作業 が必要なため、コーダの負担軽減を目的とした種々の**コーディング自動化システム**を開発してきた。大規模プロジェクトでの利用実績あり。

[問題点]

(1) 公開システムではないために一般の研究者は利用しにくい
(2) 国際比較研究の隆盛により、国際標準コードの要請も
(3) コーダの負担軽減の問題がさらに深刻化

・研究者の誰もが自由にシステムを利用できること
・国内標準コード、国際標準コードを自由に選べること
・コーダによる再コードの要／不要がわかること

研究の目的

職業・産業コーディング自動化システムを整理・統合し Webにより公開

◎ 職業・産業に関する情報のファイル(CSV形式)を**アップロード**

◎ 希望する職業・産業コード(**確信度付き**)の結果ファイル(CSV形式)を**ダウンロード**

職業・産業コードの種類

	職業コード	産業コード
国内標準	SSM職業小分類(約200個)	SSM産業大分類(約20個)
	例 501 自然科学研究者 502 人文科学研究者	例 10 農業 20 林業 30 漁業 40 鉱業
国際標準	ISCO小分類(約400個)	ISIC亜大分類(約60個)
	例 1141 Senior official of political party organizations	例 011 Growing of crops ; market gardening; horticulture

ISCO : International Standard Classification of Occupations
ISIC : International Standard Industrial Classification of All Economic Activities

自動化システム

コードの種類	前処理	自動化システムで適用される手法	後処理
SSM職業コード	形態素解析	ルールベース手法 ＋ 機械学習(SVM)	表形式変換
SSM産業コード		ルールベース手法	
ISCO		ルールベース手法 ＋ 機械学習(SVM)	
ISIC(予定)		ルールベース手法 ＋ 機械学習(SVM)	

形態素解析はjuman (京都大学長尾研究室開発)を利用

アップロードファイルの例

NO.	学歴	地位＆役職	事業規模	勤務先事業の種類	仕事の内容
1	9	9	8	工場	コピー機のトナーカートリッジの製造
2	9	3	6	工場	ガラス吹き
3	11	4	9	福祉事務所	生活保護業務の現業員
4	11	8	8	予備校	事務
5	10	2	4	病院	看護士
	選択肢				自由記述

ダウンロードファイルの例

NO.	確信度	rank1	rank2	rank3
1	C	630	631	644
2	B	625	626	689
3	B	554	538	629
4	A	554	560	558
5	A	514	516	688
		第1候補	第2候補	第3候補

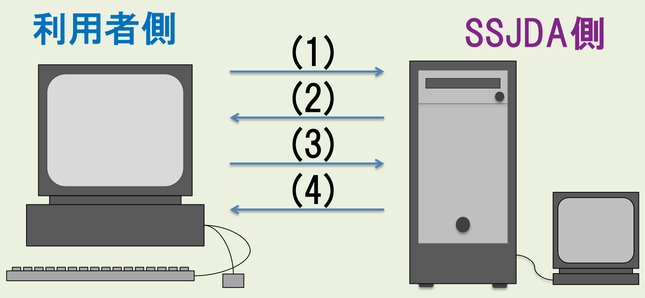
(SSM職業コードの場合)

確信度

A : コーダの作業省略可能
B : できればコーダの作業必要
C : コーダの作業必要

A: 第1位のスコア>0　第2位のスコア<=0
　第1位のスコアー第2位のスコア>α
B: 第1位のスコア>0　第2位のスコア<=0
　第1位のスコアー第2位のスコア<= α
C: A、B以外　　α は閾値

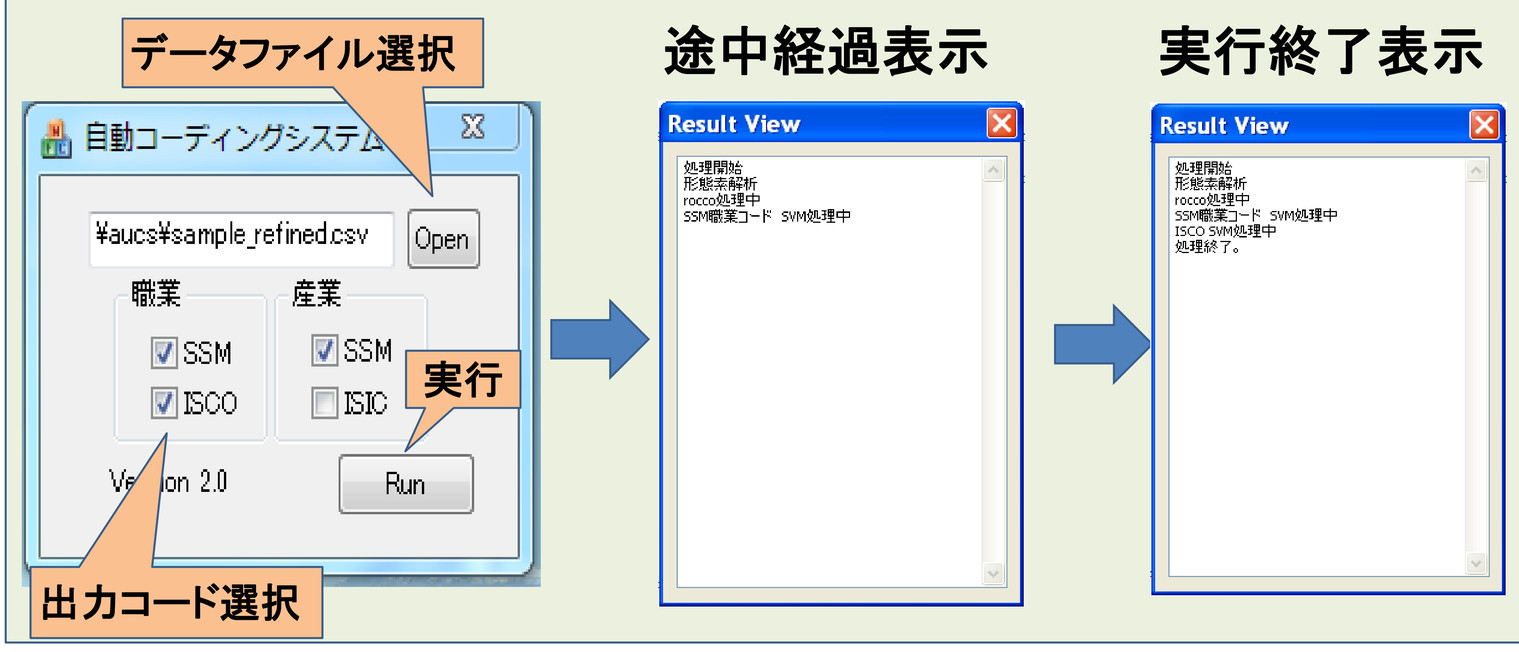
利用手順



(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 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

SSJDA側におけるオペレータ操作画面

(例)出力コード:SSM職業コード・SSM産業コード・ISCO



正解率(第3位まで)

単位:%

コード	JGSS06	JGSS08	JGSS10	SSM
SSM職業	78.8	78.9	78.3	80.6
SSM産業	70.9	77.6	73.9	70.1
ISCO	72.2	72.1	69.1	ー

確信度別正解率(第3位まで)(カッコ内はカバー率)

単位:%

コード	A	B	C
SSM職業	95.4(29)	71.6(48)	35.5(23)
ISCO	94.0(7)	67.7(67)	28.4(26)

α=3の場合

＊SSM職業コードの結果は左記データセットの平均

今後の課題

◎ システムの本格稼働
◎ ルール辞書・シソーラスの更新
◎ 確信度Aのカバー率向上
◎ 処理時間の短縮
◎ ISIC自動コーディングの組込み

連絡先:高橋　和子(敬愛大学国際学部) takak@u-keiai.ac.jp
本研究は、平成22年度～平成24年度科研費(基盤研究(C)22530516)による成果の一部である