格フレームによる職業コーディングの自動化支援システム

高橋　和子
敬愛大学国際学部
PXK10076@nifty.ne.jp

1．はじめに

社会学において、「職業」は社会的地位を測る重要な属性である。従って、社会調査では被調査者に職業を提示して選択させることをせず、仕事の内容を自由回答で収集し、これを含めた職業に関する一連のデータを分析者が総合的に判断して妥当な職業に分類するのが一般的である［5］。階層移動研究［9 他多数］のための定期的な全国調査であるＳＳＭ調査（Social Stratification and Social Mobility Survey）においても、「本人の仕事内容」を中心に収集した6種類の職業データ[1]を約 200 種類の職業コードに変換する必要があり、この作業をＳＳＭ職業コーディング（以下、職業コーディングと略す）と呼ぶ。職業コーディングはすべての分析に先立って、分析者全員の手により行われるが[2]、作業が煩雑かつ膨大なこと、コーディング結果の一貫性が保証されにくいという問題を抱えている。これは、職業データの個数が数万個にもなること[3]が原因であるが、より根本的には、コーディングの中心となる「本人の仕事内容」が自由回答であり、カテゴリーである職業の個数や定義内容が人間（コーダー）が完全に記憶できないほど複雑で多岐にわたるためである。また、多人数による長期間の作業過程で判断に揺れが生じやすいことも原因となる。

1995 年調査からは、マニュアルを繰る代わりにコンピュータでキーワード検索を行うデータベースシステム［10］が使用されたが、マニュアルに出現する語を熟知していない限り何をキーワードとすればよいかがわからない点で、少数の専門家しか使いこなせない。また、基本的に人手によるコーディングであり、一貫性の問題や効率性は改善されていないという状況である。これらの問題を解決するには、人手による単純なキーワード検索ではなく、コンピュータに回答やカテゴリーの意味を理解させて、コーディングを自動化する必要がある。本研究では、格フレームの形式による簡単な意味解析を行って、職業データを妥当な職業にコーディングするシステムを提案し評価を行う。

一方で、自由回答はテキスト情報であり、職業コーディングをテキストの自動分類であると捉えることも可能である。しかし、これまでの研究［3 他］を見る限り、次のような相違点がある。まず、テキストの自動分類は情報検索の準備として位置づけられることが多く［11］、精度より再現率が重視される傾向がある。次に、ベクトル空間モデルを用いるものが多いが、今回は 1 サンプル中に同じ単語が複数個出現するだけの長さを持たない。さらに、分類カテゴリーの範囲の広さや個数の点でも異なっている。

2．システム構築のための戦略

一般に、コーダーが自由回答のコーディングを行う際には、次の2つの処理過程を経る。

（1）回答のもつ意味内容を理解する。
（2）妥当なカテゴリーに分類してそのコードを付ける（狭義のコーディング）。

ただし、前提条件として、コーダーは、（0）カテゴリーの定義内容を知っている。すなわちカテゴリーに関する知識をもっている必要がある。

本研究では、格フレームの形式を用いて回答の意味とカテゴリーの定義内容の表現を行い（(1)、(0)）、両者を比較する際に、シソーラスによる拡張を行う（(2)）こととする。

2．1　回答の意味表現

1995 年調査の回答（約 1000 サンプル）の傾向を分析した結果、処理の対象を括弧の部分を除いた本文のみとすると、1語（34 %）のものを含めて1文が 100 %を占めた。回答の意味表現を格フレームにより行うことが可能であると

1　「従業上の地位」（選択）、「従業先の名前」（自由）、「従業先事業の種類」（自由）、「従業員数」（選択）、「本人の仕事内容」（自由）、「役職名」（自由、選択）である。
2　1985 年調査では、20 ～ 30 人×7日を費やした［*］。
3　調査規模が約 6500 サンプル（1995 年調査）と大きい上に、1つの調査票で複数の職業が尋ねられる。

判断できるものは78％であった。回答の中には、仕事の内容ではなく、職場名や役職名によるものもあるが、「課」や「係」、また「等」や「一般」など回答の意味表現に直接関係がないと思われる語を除去すると、回答の末尾にある語を格フレームの述語（動詞、サ変名詞など）としてよいものが96％あった。述語が取る格の種類は、なしが最も多くて52％、対象格が45％、場所格が2％であり、回答の意味は格フレームにより表現できるとした。

「レタスを作る」場合　　「中学校教員」場合
述語　：作る　　述語　：教員
対象格：レタス　　場所格：中学校

図1　格フレームによる回答の表現例（表層格を省略）

2．2　職業に関する知識表現

職業に関する定義は［1］［2］に記述されているが、基本的に、職業は「作る」や「教える」など動作（動詞やサ変名詞などの述語により表現される）の違いにより大まかに分類され、さらに、「何を」や「どこで」など動作の対象や動作を行う場所（名詞により表現される）により細分類される。従って、職業に関する知識も格フレームの形式により表現できる。職業ごとに知識をまとめて「職業小分類辞書」を構築する[1]。

職業コード　述語　格の種類　名詞（格スロット値）
　↓　　↓　　↓　　↓
（　５９９　栽培　（を　野菜　穀物　果樹　））
（　５９９　飼育　（を　蚕　））
（　５２２　教える　（で　中学校　））

格の種類において、「を」は対象格、「で」は場所格を表す。
名詞（格スロット値）　は格が要求する名詞を表す。

図2　「職業小分類辞書」

2．3　シソーラスの必要性

回答からカテゴリーを検索するには、両者の格フレームの内容を比較すればよいが、出現する述語や名詞に対してそれぞれシソーラスを作成して、意味的に同一視できる語をまとめたり、語の抽象度レベルの違いや表記の揺れを吸収する必要がある。

2．3．1述語シソーラス

職業を表現する述語は日常語であるために、『分類語彙表』［6］が利用できる。「述語シソーラス」（図3）における述語コードとは、『分類語彙表』においてすべての語に付いている分類番号の小数部と、グループ番号を組み合わせたものである。『分類語彙表』を利用する利点は、意味的に類似する語は品詞（「体」や「用」）に関係なく類似した分類番号が付けられていること、語に付いているふりがなも利用すれば、表記の揺れを吸収できることである。

述語　述語　述語コード
↓　↓　↓
（せいぞう　製造　３８６　１）
（せいする　製する　３８６　１）
（つくる　作る　３８６　１）

図3　述語シソーラス

なお、述語コードを用いることから、「職業小分類辞書」の述語も述語コードに直し、最終的にはこれを述語コード別に整理した「述語コード別職業小分類辞書」を作成する。

2．3．2　名詞シソーラス

「名詞シソーラス」では、「述語コード別職業小分類辞書」と回答の格スロット値をそれぞれ代表語と用例として対応付け、語の抽象度レベルの違いを解消する。職業を表現する名詞は特殊なものもあるため、［1］に基づいて構築する。

代表語　　用例
↓　　↓
（野菜　　レタス　キャベツ　きゃべつ）

図4　「名詞シソーラス」

3．システムの実現

3．1　システムの構成

システムの構成と全体の処理手順を図5に示す。システムの中心はコーディングプログラムで、形態素解析済みの職業データに対して1サンプルずつコーディングを行う。そこで参照される辞書やシソーラスは、既存のワープロ用ソフトや表計算用ソフトを利用して作成する。なお、システムで決定できなかった職業データについては、人手によりコーディングを行う。

3．2　コーディングプログラムの概要

（1）コーディングの準備（回答の編集）

①回答から括弧内の部分を除去し、本文だけに

1　実際には、この他の知識が必要な場合もあるが［2］、辞書の記述形式が煩雑になることを避けるために、プログラムで対応する（5．2（2）④'を参照のこと）。

する。

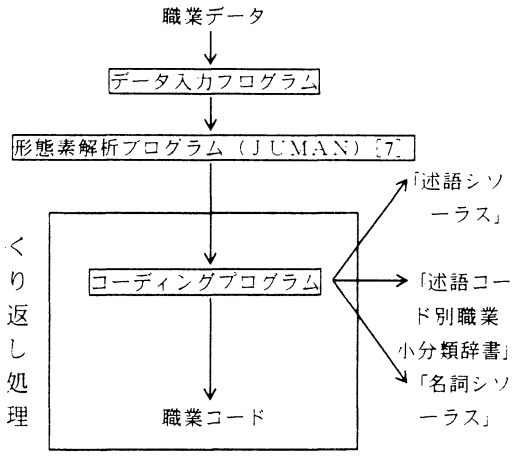

図５　システムの構成と全体の処理手順

②本文において意味解析に不要な語を除去する。
③並列表現があれば、それぞれを独立させて複数の文を作り出す。
④格助詞が省略されているものに対しては、対象格と場所格の両方を表すものとして、便宜的に「をで」なる助詞を補う。
（２）コーディング
①回答を格フレームの形式により意味表現する。
②述語を取り出して、「述語シソーラス」を検索し述語コードを付ける。検索できなければ、述語コードを「0」とし、回答の職業コードを「999」（未決定）として⑦に飛ぶ。
③述語コードにより「述語コード別職業小分類辞書」を検索して、マッチした述語コードの項にある職業コードが要求する格スロット値が、回答の名詞にあるかどうかを調べる。その際、格スロット値そのものがない場合には「名詞シソーラス」により用例を調べていく。
④回答に要求された格スロット値（または用例）とマッチする名詞があれば、その職業コードを付ける。
⑤回答にない場合は、「述語コード別職業小分類辞書」における述語コードの項にある次の職業コードに対して③④を繰り返す。
⑥格スロット値（または用例）がマッチしないまま述語コードの項にある職業コードが尽きたら、職業コードを「999」（未決定）とする。
⑦並列表現による複数の文があれば、文がなくなるまで①から⑥を繰り返す。
⑧サンプルが終了するまで、①～⑦を繰り返す。

コーディングは、基本的には上記①から⑧の手順で行われるが、より人間に近い振る舞いとして、途中で次のような処理も行う。
③’回答に、「述語コード別職業小分類辞書」により要求された格自体がないか、またはあっても特定化できない名詞（部品、製品など）の場合には、対象格であれば「従業先事業の種類」（ただし、述語が一致する場合に限る）、場所格であれば「従業先の名前」における名詞を該当する格の名詞として扱う。
④’付けられた職業コードはそのまま決定とせずに、管理職や自営業関係のチェックを行う。例えば、管理職は「従業上の地位」、「従業員数」、「役職」も参照する。

４　結果と考察

実験に用いたデータは、1995 年調査Ａ票の地区番号が００１～１１５の約 900 サンプルのうち、問４の質問ｅ「本人の仕事内容」に対して回答があったもの（無職と学生以外の 633 サンプル）である。

４．１　精度と再現率

データは全く事前編集を行わずに、調査票の回答通りに入力したものを用いた。従って、1 語中に漢字、平仮名やカタカナが混在していたり（「事ム」、「全ぱん」）、並列表現の書き方がおかしいもの（「イカ、生処理」）もそのまま処理の対象とした。ただし、明らかな入力ミス 4 個は除き、Ｎ＝629 サンプルとした。

まず、形態素解析における失敗は 40 個（6.4％）あった。そのうち、解析自身の失敗は 25 個（4.0％）、結果の表示を失敗したものは 15 個（2.4％）である。前者の原因は主に次の２つである。１つは、1 語中に異なる文字種が混在しているものが字種の違いにより別々の語に切り離されてしまったこと（「事ム」は「事　ム」と解析される。7 個）。もう１つは、例えば「米づくり」（または「米作り」）の「づくり」が、「作る」という動詞ではなく名詞性名詞接尾として解釈されたために、格フレームで表現ができなかったことである（11 個）。後者の原因は、今回、ＪＵＭＡＮの処理をすべてデフォルト値で行ったために、1 語につき 10 バイトを越えた長い語が途中で打ち切られてしまったことによる（「ウ

ェイトレス」が「ウェイトレ」と表示される)。これらの改善は容易である。

本システムに対する評価の基準は、1995 年調査後に人手で行われたコーディング結果を正解とした。その結果、「ａ正しいもの」は 390 個 (62.0 %)、「ｂ間違えたもの」は 74 個 (11.8 %)、「ｃ決定できなかったもの (職業コードが 999)」は 165 個 (26.2 %) であった。従って、精度を ａ／(ａ＋ｂ)、再現率を ａ／Ｎ として計算すると、それぞれ 84.1 %、62.0 %となった。

これを形態素解析を成功したものｎ＝589 (＝Ｎ－40) でみると、ａは 390 個、ｂ'は 71 個、ｃ'は 128 個で、精度は 84.6 %、再現率は 66.2 %である。再現率については、そもそも格フレームで意味表現が可能であると判断できたものが約 80 %程度だったことを考慮すると、その中の約 85 %が正しくコーディングされたことになる。

４．２　間違えたもの

最終的には、階層移動研究においては、職業は大分類レベル (15 個) 程度の変数として扱われることを考慮すると、間違え方を大分類が同じかどうか区別する必要がある。大分類が同じものは 21 個 (28.4 %)、間違えたもの[1]は 53 個 (71.6 %) で全サンプルの約 8 % (＝ 53 ／ 629) に相当する。間違えたものの中から大分類が同じものと他の知識を必要とする管理職や自営に関係するものを除くと、ｎ"＝ 541 (＝ｎ－(21 ＋ 27))、ａ＝ 390 個、ｂ"＝ 26 個となり、精度は 93.8 %、再現率は 72.1 %となる。

４．３　決定できなかったもの

決定できなかった原因は、先に述べた形態素解析の失敗 (40 個) と、シソーラスの不備や回答の情報不足である。シソーラスについては、今後システムを使い込むことで充実させることができるが、新語への対応を考えるとゼロとすることは不可能である。回答の情報不足については、逆にシステムの側からどのような場合にどのような情報が必要なのかを提示することが有効であると思われる。

1　最も個数が多かったのは管理にコーディングされたもの (17 個) であるが、管理の場合は、３．２ (２) ④' に示すように他の知識も必要である。

５．おわりに

本研究では、これまで人手で行っていたＳＳＭ職業コーディングを格フレームの考え方を用いて自動化するシステムを提案した。システムを実際のデータに適用した結果、クローズドテストではあるが、精度約 80 ～ 90 %、再現率約 60 ～ 70 %で有効性を示した。今後の課題は、実用化に向けてシステムを精緻化すること、ＳＳＭ以外の職業データにも対応できるようにすることである[2]。具体的には、シソーラスや辞書の充実と入力部分を含めたプログラムの改良である。

参考文献

[1] 1995 年ＳＳＭ調査研究会. 1995.『ＳＳＭ産業分類・職業分類 (９５年版)』.

[2] 1995 年ＳＳＭ調査研究会. 1995.『ＳＳＭ調査　コード・ブック』.

[3] 藤井洋一他. 1997.「共起情報を利用した文書の自動分類」.『情報処理学会研究報告』. Vol.97 No.29. 97-104.

[4] 原純輔. 1993.『ＳＳＭ職業分類 (改訂版)』.

[5] 原純輔・海野道郎. 1984.『社会調査演習』. 東大出版会.

[6] 国立国語研究所. 1964.『分類語彙表』. 秀英出版社.

[7] 松本裕治他. 1996.『日本語形態素解析システムＪＵＭＡＮ使用説明書 version3. 0』. 奈良先端科学技術大学院大学情報科学研究科松本研究室.

[8] 松本裕治. 1998.「意味と計算」.『言語の科学４　意味』. 岩波書店. 125-169.

[9] 直井優・盛山和夫. 1990.『現代日本の階層構造①社会階層の構造と過程』. 東大出版会.

[10] 佐藤嘉倫. 1992.「職業コーディング支援システムの構築」. 原純輔 (編)『非定型データの処理・分析法に関する基礎的研究』. 平成 3 年度文部省科研費補助金 (総合Ａ) 研究成果報告書. 199-204.

[11] 佐藤理史. 1996.「情報の構造化と検索」.『自然言語処理』(長尾　真編). 岩波書店. 412-457.

[12] 高橋和子.2000.「自然言語処理に基づく自由回答のコーディング支援－格フレームによるＳＳＭ職業コーディング自動化システム」『数理社会学会論文誌　理論と方法　２７』. Vol.15　No.1. (予定)

[13] 徳永健伸. 1999.『情報検索と言語処理』. 東大出版会.

[14] 安田三郎・原　純輔. 1982.『社会調査ハンドブック第３版』. 有斐閣.

2　現在、東京大学社会科学研究所の調査における職業コーディングに向けて改良中である。