職業コーディングにおけるROCCOシステムとSVMの組み合わせ

高　橋　和　子

（敬愛大学国際学部）

A combination of ROCCO system and Support Vector Machines in occupation coding

Kazuko TAKAHASHI

We have used Rule-based OCcupation COding (ROCCO) system, which is a rule-based automatic method, to the occupational/industrial data of JGSS-2000/2001/2002. The performance of ROCCO system is more stable than that of coders and by using it, we can lighten work of coders and save much time. However, its categorization performance is not satisfiable. Therefore, we adopt Support Vector Machines (SVMs) , which show high performance in various fields, and compare it with the rule-based method. We also investigate effective combination methods of SVMs and ROCCO system, especially the method we use the results of SVMs if ROCCO can't decide occupational code. We empirically show that SVMs outperform ROCCO system in the occupation coding and that the combination of the two methods yields an even better accuracy. We can improve the accuracy by contriving the combination of two methods.

Key words: Occupation Coding, Rule-based method, Support Vector Machines

JGSSにおいては、ルールベース手法により職業コーディングを自動的に行うシステム（Rule-based Occupation Coding；ROCCO システム）が利用されてきた。ROCCO システムはコーダーと比較すると性能が安定しており、コーダーの負担を軽減する。しかし、正解率に注目すると高いとはいえず、またルールベース手法に固有な問題から現在の値以上にすることは困難であると思われる。そこで、 機械学習の一つで分類性能が高いとされるサポートベクターマシン（Support Vector Machine, SVM）を適用し、ROCCOシステムとの比較を行った。さらに２つの手法の組み合わせ方として、特にROCCOシステムが職業コードを決定できない場合にSVMの結果を利用する方法についての検討を行った。その結果、SVMはROCCOシステムより正解率が高く、２つの手法を組み合わせることでさらに正解率を高めることができた。組み合わせ方を慎重に行うことでさらに正解率を高めることが可能である。

キーワード：職業コーディング、ルールベース手法、サポートベクターマシン

１．はじめに

職業・産業コーディングとは、社会調査において自由回答により収集される職業や産業に関するデータを該当する職業コードや産業コードに分類する作業をいう。調査によっては、1 つのサンプルに対して、本人の現職だけでなく、初職から現在までの職歴、配偶者の職業、父親の職業など多数のコーディングを行う必要がある上に、職業の場合は通常、小分類まで行う関係上カテゴリ数が多い（約 200 個）ため、コーダーの負担が非常に大きいという問題がある。また、コーディングの作業量が膨大なために多人数で長期間作業を行うことの弊害として、コーディングの結果に一貫性が欠けるという問題も生じている。

JGSS ではこの点を考慮し、第１回本調査（JGSS-2000）の時点から、コンピュータにより職業･産業コーディングを自動的に行うシステム(Rule-based OCcupation COding;ROCCO)（高橋, 2000）を利用してきた。人間とコンピュータの組み合わせによるコーディングはこれまでに例がなく、実際に ROCCO システムを利用しながら有効な併用方法を模索してきたが、次に述べるような方法で実現することとした（詳細については西村・石田 2001 を参照のこと)。まず、ROCCO システムとコーダーが別々にコーディングを行う。次に、コーダーは ROCCO システムの結果を参考にしながら見直しを行い、必要ならば訂正を行う。さらに、これらすべてを専門家が見直し、判断に迷う場合には複数の専門家による協議を行って、最終的な決定を行う。

ROCCO システムの利用により、JGSS では、コーダーが行うべき作業の１回分をコンピュータが担当することになり、コーダーの負担が軽減された。また、ROCCO システムの出力結果はコーダーが判断する際のヒントとして使えるため、コーディング作業の支援という効果もあった(1)。

ROCCO システムの性能は、過去 2 回行われた JGSS-2000 と JGSS-2001 のいずれでも安定しており、平均で職業においては精度約 80%、再現率約 66%、産業では精度約 93%、再現率約 75%であった（高橋, 2003)。コーダー(2)と比較すると、職業コーディングの精度は JGSS-2000 では 1～8%程度上回っていたが､JGSS-2001 では同じ程度下回った｡これは、両年ともほぼ同じメンバーであったコーダーの「習熟」度ほど、ROCCO システムで行われた辞書（職業を決定するためのルールをまとめたもの）やシソーラス（用語を拡張するために類義語や同義語をまとめたもの）のメンテナンスの程度が及ばなかったものと思われる。

実は、ここに ROCCO システムを始めとするルールベース手法の限界がある。ルールベース手法ではルールにマッチしないデータは処理できないが、職業・産業コーディングにおけるすべてのルールをあらかじめ用意するのは非常に困難な問題なのである。例えば、管理職関係の職業に対する判断を行うための原則(3)は存在するが、実際には必ずしも原則通りに決定されるわけではなく、これらをルール化することは不可能に近いといえよう。さらに、職業や産業は処理の対象となるデータが言語であるために、新しい用語の出現や使

い方が次々と出現する。従って、メンテナンスを半永久的に続けていかなくてはならず、労力の面でも大きな問題となる(4)。

これに対して、最近、多方面で利用が進んでいる機械学習では、コンピュータに訓練用の事例を与えて学習させて、新たな事例を判断させるために、ルールベース手法で必要となるルールの作成やメンテナンスを行わずにすみ、また、手続き処理を記述するプログラムを作成する必要がないという利点をもつ。そこで、次回（JGSS-2003）の職業・産業コーディングに向けて、新たに機械学習の適用も検討することとした。その中で、自然言語処理において分類性能のよさで注目されているサポートベクターマシン（Support Vector Machine；SVM）（工藤, 2002）を JGSS-2001 における職業データに対して実験した結果、有効性が示唆されたため（高橋他, 2003）、実用に向けての実験を重ねてきた。その結果、SVM はルールベース手法より正解率（再現率）(5)が高いこと、また、ROCCO システムと組み合わせること(6)でさらに高くなることが確認できた（高橋他, 2004a 2004b）。

本稿における主要な目的は、この 2 つの手法による組み合わせ方の一つで、「ROCCO システムが職業コードを決定できない場合に SVM の結果を利用する方法」に関するさらなる検討を行うため、ROCCO システムが決定できない場合に利用される SVM の結果についての状況を調査することである。

以下では、次節でまず、前回に引き続き基本的な報告として、JGSS-2002 における職業・産業コーディングの結果を述べる。次に、3 節で職業コーディングに SVM を適用する方法と結果について述べ、4 節で SVM と ROCCO システムを組み合わせる方法と結果を、本稿に関連するものに限定して述べる。5 節では、ROCCO システムが決定できない場合に利用される SVM の結果について、その正誤状況を中心に考察する。最後にまとめと今後の課題を述べる。

２．ROCCO システムによる職業・産業コーディング（JGSS-2002）の結果

JGSS-2002 におけるサンプル数は 2,958 で、コーディングの対象は、これまでと同様に、本人現職、本人最後職、本人初職、配偶者職、父職の 5 種類である（ただし、父職は産業データが収集されないために 4 種類）。

ROCCO システムによる結果を表１、表２に示す。ただし、今回は、SVM による方法との比較を考え、複数個出力した場合は最初の１個（プログラムの都合上、回答においては最後に記述されたもの）しか見ていないために、これまでの性能算出方法と比較すると不利な条件となっている(7)。結果は、5 種類の職業（産業の場合は 4 種類）の平均で、職業の精度が 79.4%、同再現率が 67.7%、産業の精度が 81.0%、同再現率が 74.7%であった。

過去 2 年間の結果と比較すると、職業では、平均で精度が 0.9%下がり、再現率が 1.5%上がっているが、いずれも小さな変動で前回と同様に安定した結果を示している。5 種類の職業の中で精度が前回より上がったのは本人最後職と父職で、それ以外の 3 種類のもの

は下がっており、本人現職が最も悪いのはこれまでと同様である。ただし、再現率は配偶者職以外のものはすべて上がっている。先に述べた条件を考慮すると、全体的に性能はよくなっている。

産業は、JGSS-2001 はすべての種類において JGSS-2000 を上回っていたが、今回はすべて JGS-2001 だけでなく KGSS-2000 をも下回った。これまでの結果と比較すると、平均で精度が 12.1%、再現率が 0.4%下がっている。これは、前述した条件のためだけでなく、今回初めて、産業辞書やシソーラスに対するメンテナンスを全く行わなかったことが原因であると思われる。このように、言語である自由回答が中心となる職業や産業データを対象とするシステムにおいては、シソーラスの恒常的なメンテナンスが必要であることが再確認できた。

表１　ROCCO システムによる職業の精度と再現率（単位：%）

本人現職		本人最後職		本人初職		配偶者職		父職	
精度	再現率	精度	再現率	精度	再現率	精度	再現率	精度	再現率
76.8	63.3	80.5	67.7	77.0	68.3	77.4	64.6	85.3	74.6

表２　ROCCO システムによる産業の精度と再現率（単位：%）

本人現職		本人最後職		本人初職		配偶者職	
精度	再現率	精度	再現率	精度	再現率	精度	再現率
79.5	72.3	82.4	75.8	83.3	78.1	78.7	72.6

３．SVM による方法

SVM は統計的機械学習に基づく 2 クラスのパターン認識手法である。学習サンプルと分類境界の間隔（マージン）を最大化するという戦略により、ニューラルネットワークなどの従来手法より、汎化能力が高く、精度よく評価事例を分離できるという特徴をもつ。このような優位性から、SVM は、自然言語処理の分野においても、文書分類や係り受け解析への応用で高い性能を示す。

SVM による方法では、ルールだけでなく手続きを記述するプログラムを作成する必要もない。これは手間がかからないだけでなく、分野や対象に依存しないという汎用性の点でも評価できる。ただし、分類の性能を上げるために重要になる素性の選び方は、実験から得られた結果により判断されるため、試行錯誤的に実験を重ねる必要がある。

SVM は基本的に二値分類器であるため、今回の職業小分類のような多値分類に適用するには拡張を行う必要がある。本稿では、one-versus-rest 法を用いて多値分類器へと拡張した。SVM による方法では、分離平面からの距離がすべてマイナスの値の場合（負例）でも、その中で最も小さなカテゴリに分類されるため、必ず何らかの職業コードを出力する。この点が、判断に迷う場合には無理にコードを決めず、未決定のコードを出力する ROCCO

システムと異なる。

ROCCO システムとの組み合わせる際に利用される結果を得るために、今回は、基本的な素性（以下、基本素性と呼ぶ）として次のものを用いた[8]。

・「仕事の内容」に出現する単語（原形）[9]

・「従業先事業の種類」に出現する単語（原形）

・「従業上の地位＋役職」の選択肢（14 種類）

実験は実際のコーディングを想定して、過去に蓄積されたデータを訓練データ、これからコーディングを行うデータを評価データと考えた。従って、利用できるデータを、訓練データとして JGSS-2000 と JGSS-2001 の本人現職（または最後職）、本人初職、配偶者職の 4 種類のうち無職と学生を除く有職者[10]、評価データとして JGSS-2002（同）に分割した。その結果、訓練データは 13,296（＝6,848＋6,448）サンプル、評価データは 6,770 サンプルとなった。また、ソフトマージンパラメタ C[11]の値を 1.0～0.1 まで 10 通りに変化させて実験を行った。

SVM による結果を表３に示す。参考のため、ROCCO システムによる結果を C=1.0 欄の右列に示した。C の値に関係なく、SVM の方が ROCCO システムより正解率が高いことがわかる。

表３　SVM による正解率（単位：%）

	SVM	ROCCO
C=1.0	71.9	67.7
max	71.9	―
min	70.3	―

４．ROCCO システムと SVM の組み合わせ

前節で述べたように、SVM は単独で適用しても ROCCO システムより正解率が高い。しかし、ROCCO システムの利点は、決定した職業コードについては SVM より正解の割合（精度）が高いことである[12]。この点を考慮すると、SVM と ROCCO システムを組み合わせる方法を検討するのは有効である[13]と考えられる。ここでは、その中の一つで、2 つの手法の出力結果をシーケンシャルに利用する方法として、「ROCCO システムが決定できない場合に SVM の決定を利用する」方法（以下、seq と略す）について述べる。

seq では、まず ROCCO システムにより出力される職業コードを調べ、それが未決定コード「９９９」であったり、出力されたコードが複数個ある場合は、ROCCO システムが職業コードを決定できないと判断して、SVM により出力された職業コードを用いる。

ROCCO システムが SVM による方法と異なるのは、必ずしも職業コードを決定しない点である。この理由は、トレードオフの関係にある再現率と精度においては精度を優先するという設計方針により、曖昧なものは無理にコードを付けないままにしておくためである。その結果、ROCCO システムでは未決定のコード（９９９）が全体の約 20%程度[14]と多くな

る傾向がある。このとき、設計方針とは矛盾するが、ROCCO システム以外にもコードを出力してくれる装置があれば、情報がゼロである「９９９」のデータに対してコードに関する何らかの情報を与えてもらいたいという要求に対しては、SVM の結果を利用することが可能である。

seq の実験の結果を表４に示す。参考として、SVM 単独による結果と ROCCO システムによる結果も示した。seq の正解率が ROCCO システムより高いのは当然であるが、SVM 単独の場合と比較しても、C の値に関係なく常に高い。その場合、max と min の差が非常に小さく（0.2%）、SVM における差が 1.6%であることと比較すると、C の値による影響をほとんど受けないことがわかる。

表４　seq による正解率（単位：%）

	seq	SVM	ROCCO
C=1.0	72.9	71.9	67.7
max	73.1	71.9	―
min	72.9	70.3	―

表４により seq の有効性が確認できたが、そこで利用される SVM の結果における正誤の状況は全く考慮されていない。seq の正解率を高めるためには、この点に関する検討を行う必要があると考えられるため、次節で調査する。

５．　seq において利用される SVM の結果について

本稿において、ROCCO システムが職業コードを決定できないとは、次の a～c の場合をいう（高橋他，2004a）。

（a）未決定コード「９９９」を出力する

（b）形式的には１つの職業コードを出力するが、内容的に複数個含む

（例）職業コード「５５７０５７３」・・・５５７（営業・販売事務員）

または５７３（外交員（保険、不動産を除く））

（c）回答に複数個記述があるために複数個出力する

（例）回答「○○の製造と販売」　→　職業コード「７０３」「５６９」を出力

ここで、（a）となるのは、回答の情報が不足する場合（この場合はコーダーも SVM も決定するのが困難である）と、回答には十分な情報があるが ROCCO システムの辞書やシソーラスの不備により職業コードを決定できない場合の 2 通りである。seq において SVM に期待されるのは、後者の場合である。（b）は（a）の特別な場合で、「５５７」であるか「５７３」であるかが人間でも判断のつきにくい回答に対して、「９９９」にするには惜しいという判断により設けたものである。（c）は、ROCCO システムが回答の１つ１つに対して職業コードを付けるために生じる状況である。ROCCO システムにおいても、出力された職業

コードのペアから出力する職業コードを決定するルールを作成することが不可能ではないが、起こり得るすべてのルールを記述するのは困難である(15)。

ROCCO システムが決定できない場合が生じる割合(16)を職業の種類別にみると、ROCCO システムが「９９９」を出力する場合は本人現職が 11.5%で最も多いが、続く本人初職（9.5%）、本人最後職と配偶者職（いずれも 8.7%）ではそれほど差がない。また、決定できない別の 2 つの場合でも、すべての値が 1.2～2.0%にあり、ほぼ同じ値であるといえる。これより、ROCCO システムが決定できない場合は、本人現職で「９９９」がやや多く出力されることを除けば、職業の種類による偏りがないといえる。すなわち、職業の種類が異なっていても、この問題に関しては同じ性質のデータとして扱うことができる。

以下では、ROCCO システムが決定できない場合の SVM による結果の正誤について、「ROCCO システムが決定できない場合別」にみた場合と、「カテゴリ別」にみた場合の２つの観点から調査する。

5.1 ROCCO システムが決定できない場合別の状況

SVM による結果の正誤状況を、ROCCO システムが決定できない3つの場合に分けて示す（表５）。

表５より、まず、ROCCO システムが職業コードを 1 つに決定することができないのは 1,771 サンプルで、これは全サンプルの 26.2%に相当する。これらの事例に対して、SVM が正しく決定しているものは誤って決定しているものより少なく 48.6%でしかない。従って、SVM の結果をそのまま利用するだけでは効果が上がりにくいことがわかる。

これを 3 つの場合に分けて調べてみると、(a) ROCCO システムが「９９９」を出力する場合は、SVM による結果は約 55%のものが誤っており、正しい場合より約 10%程度も多い。(b) 1 つの回答に複数個の内容がある場合は約 60%のものが正しいが、(c) 回答に複数個の記述がある場合は正しいものは約 50%程度しかない。これより、特に「９９９」を出力した場合に SVM の結果を利用するときは、慎重に行う必要があることがわかる。

表５　ROCCO システムが決定できない場合に利用される SVM の結果における正誤状況

ROCCO の出力	SVM の結果が正しい個数(%)	SVM の結果が誤り個数(%)	計
(a) ９９９	５３３ (44.9)	６５５ (55.1)	１１８８
(b) 内容が複数個	１６４ (62.4)	９９ (37.5)	２６３
(c) 回答が複数個	１６４ (51.3)	１５６ (48.8)	３２０
計	８６１ (48.6)	９１０ (51.4)	１７７１

表５を表３と比較すると、正解率がそれぞれ 27.0%、9.5%、20.6%低いことから、ROCCO システムが決定できないもの、特に、「９９９」を出力する場合や回答に複数個の記述がある場合には、SVM でも正しく決定するのが困難であると考えられる。

5.2 カテゴリ別の状況

ここでは、SVM による結果の状況についてカテゴリ別に調べる。

JGSS で用いられるカテゴリは、SSM 職業分類の５０１～６８８までの 188 個に７０１～７０５の 5 個を追加した計 193 個（９９８と９９９を除く）である。このうち、(a) ROCCO システムが「９９９」を出力する場合に SVM が出力したカテゴリ数は 151 個（78.2%）で、(c) 回答に複数個の記述がある場合は 76 個（39.4%）であった。また、(b) 1 つの回答に複数個の内容がある場合は、出現する可能性のあるカテゴリはあらかじめルールに記述されており明らかであるが、実際に出現したものは 4 種類（9 個）（4.7%）であった[17]。

紙面の都合上、カテゴリごとの正誤状況表を掲載できないため、以下では、正解率に注目し、まとめて報告する。

(a) ROCCO システムが「９９９」を出力する場合に、正解率が 80%以上のものは 20 個（出現したカテゴリ中の 13.2%。以下同様）で、そのうち 16 個（10.6%）[18]は 100%の正解率であった。逆に、20%以下のものは 50 個（33.1%）で、そのうち 41 個（27.2%）は 0%であった。残り 81 個（53.6%）は 20%より高く 80%より低いものである。

同様に、(b)回答に複数個の記述がある場合に、正解率が 80%以上のものは 20 個(26.3%)で、そのうち 18 個（23.7%）[19]は 100%の正解率であった。逆に、20%以下のものは 25 個（32.9%）で、そのうち 23 個（30.1%）は 0%であった。残り 31 個（40.8%）は 20%より高く 80%より低いものである。

正解率が 100%のものは積極的に利用していけるカテゴリであると考えられるが、「９９９」が出力される場合と回答に複数個の記述がある場合で傾向が異なっており、一致するのは、「５２３」（高等学校教員）と「６６１」（塗装工、画工、看板工）の 2 個しかない。80%以上のものを含めても、「５９９」（農耕・養蚕作業者）がプラスされるだけである。また、正解率が 0%のカテゴリは利用する際に最も注意が必要であるが、これも両者で一致するものは、「５４５」(管理的公務員)、「５４８」(会社役員)、「５７２」(商品仲立人)、「６６８」（かばん・袋物製造工）の 4 個だけで、20%以下のものを含めても、「５９０」（下宿・アパートの管理人、舎監、寮母)、「６３０」(金属工作機械工、めっき工、金属加工作業者)、「６８４」(現場監督、その他の建設作業者)の 3 個がプラスされるだけである。

このように、ROCCO システムの決定できない状況により SVM が出力するカテゴリの信頼性が異なるため、SVM の結果を利用する際には、利用に至った ROCCO システムの状況と関連づけておく必要がある。

６． おわりに

本稿では、まず、JGSS-2002 の職業・産業コーディングに対する ROCCO システムの性能（精度と再現率）を報告し、前回までと併せた 3 年間の報告を通じて、ROCCO システムの

安定性を明らかにした。しかし、現在の手法では正解率の大幅な向上を期待することは困難であることから、次に、機械学習の一つであるSVMによる方法の適用を行った。その結果、ROCCOシステムより正解率が高いことがわかった。さらに、2つの手法を組み合わせるものとして、両者の結果をシーケンシャルに利用する方法を検討した。今回は、「ROCCOシステムが決定できない場合にSVMの結果を利用する方法」についての実験を行ったが、SVM単独の場合よりも正解率が高かった。そこで、この方法の有効性を高めるために、利用されるSVMによる結果について正誤の状況を調査した。その結果、全体では正解の方がやや少なかったが、どのような場合にSVMの結果が利用されるかにより、その状況が異なることがわかった。

今後の課題は、両者の結果をシーケンシャルに利用する別の方法として、「ROCCOシステムとSVMによる方法の結果が不一致の場合にどのような決定を行うか」についての検討を行うことである。

[謝辞]

本稿の作成において、東京工業大学精密工学研究所奥村学助教授および高村大也助手に多大なご協力をいただいたことを記して感謝いたします。また、ＳＳM職業分類の使用に当たっては、東北大学文学部原純輔教授に快諾していただいたことについて感謝いたします。

[Acknowledgement]

日本版General Social Surveys（JGSS）は、大阪商業大学比較地域研究所が、文部科学省から学術フロンティア推進拠点としての指定を受けて（1999-2003年度)、東京大学社会科学研究所と共同で実施している研究プロジェクトである（研究代表：谷岡一郎・仁田道夫、代表幹事：佐藤博樹・岩井紀子、事務局長：大澤美苗)。データの入手先は、東京大学社会科学研究所附属日本社会研究情報センターSSJデータ・アーカイブである。

[注]

(1) この観点を発展させて、現在、コーダーのコーディング作業そのものを積極的に支援する「タグ付け支援システム」（NANACOシステム）を開発中である（高橋他, 2004c)。

(2) JGSS-2000における職業の精度は約75.9%、再現率は約73.7%で、JGSS-2001ではそれぞれ約84.8%、約83.5%であった（高橋, 2003）

(3)『ＳＳM調査　コード・ブック』によると、管理職については次のようにコードすることになっているが、あくまで原則に過ぎない。

① 従業上の地位が役員または自営業主の場合

規模５人未満　…必ず管理的職業以外の仕事の内容でコードする。

規模３０人未満…管理的職業以外の仕事の内容を優先してコードする。

規模３０人以上…原則としていずれかの該当する管理的職業でコードするが、
それ以外の仕事の内容が書いてあれば、それに従ってコードする。

② 従業上の地位が一般従業者や家族従業者である場合

役職が課長以上　①と同様。

役職が課長補佐以下　必ず、必ず管理的職業以外の仕事の内容でコードする。

③ 専門的管理職（設計技師長、病院長、学校長など）は「専門」の方を優先する。

(4) さらに、ROCCO システムの場合は、職業の定義を格フレームの形式で表現しルール化するために、これとマッチングをとるためには回答も格フレームの形式で表現する必要があるために、この形式で表現できない回答（約 20%弱程度存在する）に対しては処理できないという欠点がある。

(5) SVM における「正解率（Accuracy)」は、これまで高橋（2000, 2001, 2002, 2003a）で使用してきた「再現率」のことである。すなわち、

正解率＝再現率＝正解の個数／全サンプル数

また、精度は、次の意味で用いている。

精度＝正解の個数／未決定（９９９）以外のコードが付いた個数

(6) SVM と ROCCO システムの組み合わせとしてこれまでに実験を行ったのは次の 3 通りである。詳細は高橋他（2004a）を参照のこと。

・SVM の素性として ROCCO システムが出力した職業コードを追加する
・SVM の素性として ROCCO システムでマッチしたルールを追加する
・ROCCO システムが決定できない場合に、SVM の結果を利用する

(7) 一般に、回答には主要なものから順に記述されることを考えると、さらに不利であることが予想される。

(8) 素性選択を含め、用いる素性をさまざまに変えた実験を行ったが（高橋他, 2004a)、紙面の都合上、本稿では説明を省略する。

(9) 形態素解析ソフト JUMAN（黒橋・長尾,1999）により切り出された語の「原形」と品詞を用いた。

(10) 無職と学生以外とする判定の仕方は、「従業上の地位＋役職」の回答が非該当以外としたため、一部に記入ミスによる有職書以外のデータが混入していた。

(11) ソフトマージンパラメタ C は SVM の例外的な事例に対する許容度を表すもので、値が小さいほど例外的な事例に与える重みが小さくなる。

(12) SVM による方法では職業コードを 1 個に絞って出力するのに対し、ROCCO システムでは例えば回答に複数個記述がある場合には、それぞれに対して出力するため複数個となる（5 節（b)、(c）を参照のこと)。その場合、性能の計算は、その中に正解が含まれていれば正解としていたために、SVM に比較すると ROCCO システムの方が高目に出ていた。これを、複数個出力したときに正解が含まれている場合にはそのまま正解とせず、出力した個数で

割った値にすること（例えば0.5など）で補正した結果をSVMと比較したが、ROCCOシステムの方が高かった（高橋他，2004a）。

(13) ROCCOシステムは機械学習手法ではないが、一種のensemble learningであると考えることもできる。

(14) JGSS-2002の場合は約25%であった。

(15) ROCCOシステムに対して、1つに絞り込まずに対応するコードをそのまま出力するという要請もある。

(16) 職業の種類別の割合は次の式により計算した。

職業の種類ごとのROCCOシステムが決定できない個数／職業の種類ごとの全サンプル数

(17) 1つの内容に複数個ある場合は次に示すようにあらかじめわかっているために、特に調査の必要がない。

5570573（557と573）　50305040506（503と504と506）　6070686（607と686）　5560685（556と685）

(18) 正解率が100%のものは次の通りである。「514」「521」「522」「523」「547」「550」「560」「574」「595」「604」「611」「614」「632」「641」「661」「681」

(19) 正解率が100%のものは次の通りである。「516」「523」「524」「531」「532」「561」「579」「633」「634」「635」「640」「651」「652」「654」「658」「661」「672」「702」

[参考文献]

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

1995年ＳＳＭ調査研究会，1995,『ＳＳＭ調査　コード・ブック』.

工藤拓，松本裕治，2002,「Support Vector Machineを用いたChunk同定」,『自然言語処理』9(5)，3-22.

黒橋禎夫・長尾真，1999,『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.

西村幸満・石田浩，2001,『JGSS-2000調査（2000年11月）　職業・産業コーディングインストラクション』，東京大学社会科学研究所.

大阪商業大学比較地域研究所，東京大学社会科学研究所（編），2003,『日本版General Social Surveys JGSS-2001　基礎集計表・コードブック』東京大学社会科学研究所.

P.Seng-Bae and Z.Byoung-Tak，2003，Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learning，*Proceedings of the 41-th Annual Meeting of the Association for Computational Linguistics(ACL2003)*，497-504.

高橋和子，2000,「自由回答のコーディング支援について―格フレームによるＳＳＭ職業コーディングシステム―」,『理論と方法』，15(1)，149-164.

高橋和子，2001,「自由回答のコーディング自動化システム―「健康と階層」調査における職

業コーディング―」,『敬愛大学国際研究』, 8.31-52.

高橋和子, 2002,「職業・産業コーディング自動化システムの活用」,『言語処理学会第 8 回年次大会発表論文集』, 491-494.

高橋和子, 2003a,「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-2001 で見た日本人の意識と行動』, 大阪商業大学比較地域研究所・東京大学社会科学研究所（編), 179-192.

高橋和子, 高村大也, 奥村学, 2003b,「機械学習による職業コーディング―ルールによる自動コーディングシステムとの比較―」,『第 36 回数理社会学会大会研究報告要旨集』, 68-71.

高橋和子, 高村大也, 奥村学, 2004a,「機械学習とルールベースによる職業コーディング」, 情報処理学会第 159 回自然言語処理研究回報告（予定).

高橋和子, 高村大也, 奥村学, 2004b,「機械学習とルールベースの組み合わせによる職業コーディング」,言語処理学会第 10 回年次大会発表報告（予定).

高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学, 2004c,「タグ付け支援システムの開発―職業コーディング支援システム―」,第 37 回数理社会学会大会研究報告（予定).