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

Kazuko Takahashi[1], Hirofumi Taki[2], Shunsuke Tanabe[3] and Wei Li[4]
[1]*Faculty of International Studies, Keiai University, Inage-ku, Chiba-city, Japan*
[2]*Faculty of Social Science, Hosei University, Machida-city, Tokyo, Japan*
[3]*Faculty of Letters, Arts and Science, Waseda University, Shinjuku-ku, Tokyo, Japan*
[4]*Graduate School of Science and Engineering, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan*

Keywords: Automatic Coding System, Answers to Open-Ended Question, Occupation and Industry Coding, Natural Language Processing, Machine Learning, Confidence Level.

Abstract: We develop a new automatic coding system with a three-grade confidence level corresponding to each of the national/international standard code sets for answers to open-ended questions regarding to respondent's occupation and industry in social surveys including a national census. The "occupation and industry coding" is a necessary task for statistical processing. However, this task requires a great deal of labor and time-consuming. In addition, inconsistent results occur if the coders are not experts of coding. In formal research, various automatic coding systems have been developed, which are incomplete and generally unfriendly to a non-developer user. Our new system assigns three candidate codes to an answer for coders by SVMs (Support Vector Machines), and attaches a three-grade confidence level to the first-ranked predicted code by using classification scores to support a manual check of the results. The system is now open to the public through the Website of the Social Science Japan Data Archive (SSJDA). After the submitted data file which followed the specified format is approved, the users can obtain files of codes for up to four kinds with a three-grade confidence level. In this paper, we describe our system and evaluate it.

1 INTRODUCTION

We propose an automatic coding system with a three-grade confidence level corresponding to each of the national/international standard code sets for answers to open-ended questions (free answers) regarding a respondent's occupation and industry in social surveys including a national census.

When data samples on occupation and industry are collected mainly as free answers, researchers need to assign one of nearly 200 occupation codes (20 industry codes) to each sample before data analysis, which is called "occupation and industry coding" (Hara, 1984). However, because of the many kinds of classified codes (1995SSM Survey Research Group, 1995), the complexity of coding rules (1995SSM Survey Research Group, 1996) and the large quantity of samples, this task requires a great deal of labor and time-consuming (Seiyama, 2004). In addition, if the coders are not experts of the coding, inconsistent results occur with high possibility (Todoroki and Sugino, 2013).

In formal research, various automatic coding systems have been developed to support coders (Takahashi 2000; Takahashi et al., 2005a, 2005b), and are used especially in large-scale surveys such as JGSS (Japanese General Social Surveys) (Takahashi, 2002, 2003; Takahashi et al., 2005c) and SSM (Social Stratification and social Mobility) surveys. However, these systems can still be improved. First, when a researcher requests the international standard codes for international comparative studies, there are no systems can assign them yet. Second, these systems can not indicate the reliability of automatic coding results although coders desire to know them. Third, these systems are normally not open systems and unfriendly for users.

In this work, we have developed a new integrated system corresponding to each of the

Takahashi K., Taki H., Tanabe S. and Li W..
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.
DOI: 10.5220/0005131703690375
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2014), pages 369-375
ISBN: 978-989-758-049-9
Copyright © 2014 SCITEPRESS (Science and Technology Publications, Lda.)

national/international standard code sets. The new system assigns three proper codes from the first-ranked to the third-ranked to an answer by applying the methods of SVMs (Vapnik, 1998) such as an effective combination method of SVMs and the rule-based method. We have also added new functions to the system attaching a three-grade confidence level to each predicted code by using classification scores to support a manual check of the results. The levels A, B and C represent a high degree, middle degree and low degree, respectively. Furthermore, a Web-based general release is developed, considering it more effective and convenient for researchers. The public can access the system through two methods. One is that we expose the system itself and a user runs the software by downloading it (SOIC, 2000). The other method is that a user can obtains results by downloading them after uploading data on the Website without exposing the system (NIOCCS, 2013). We have adopted the latter method for the reason that the first method is difficult for the user because the user must perform maintenance on the rather complex operating environment of the system. In terms of maintenance issues and network security, we have determined that the system should be used off-line by an operator even though this method may be somewhat antiquated.

The proposed system is now open to the public by the Website of the Social Science Japan Data Archive (SSJDA) (SSJDA, 2013). When a user sends a CSV data file in a specified format on the Website of SSJDA after approval, he/she can obtain CSV result files of codes for up to four kinds with a confidence level for each code.

This paper is organized as follows. In Section 2 we discuss related work. In Section 3 we describe the proposed system. In Section 4 we evaluate the system. Finally, we conclude the paper in Section 5.

2 RELATED WORK

In Korea, there exists a Web-based AIOCS (A Web-based Automated System for Industry and Occupation Coding) (Jung et al., 2008). When a user enters a content company name (free answer), business category, department, job title, the work (free answer) on the Website, the result is displayed on the same screen later. Jung et al. (2008) reject SVMs because they need a large memory and CPU time requirements when a computer is trained on large samples with more than 400 classes in their case. When the system is adopted in the order of a rule-based method, Maximum Entropy Method and Information Retrieval Technology, its precision is the best (76%).

In the United States, the SOIC (Standardized Occupation & Industry Coding) software may be downloaded free of charge on the Web of the CDC (Centers for Disease Control and Prevention) (SOIC, 2002), which assigns codes by mainly matching the rules according to the 1990 Census. The Assigned codes that matched manually assigned codes are 75% (occupation codes), 76% (industry codes), 63% (both occupation and industry codes), respectively.

NIOCCS (The NIOSH Industry & Occupation Computerized Coding System) (NIOCCS, 2013) is a successor of SOIC, which assigns codes according to the 2000 Census. NIOCCS also assigns codes by mainly matching rules. NIOCCS has a single record coding mode in which a user enters their Age, Education, Company (free answer), City of Employment (free answer), State of Employment, Employment Zip code, and a batch file coding mode. In both modes, NIOCCS sets the three confidence levels (High, Medium and Low) to the results.

3 PROPOSED SYSTEM

In this work, the proposed system can assign four kinds of codes to a sample as follows. As a national standard code set based on a national census (1995SSM Survey Research Group, 1995), (a) Unit groups of SSM occupation code set and (b) Major groups of SSM industry code set are assigned. As the international standard code set defined by ILO, (c) Unit groups of an ISCO (International Standard Classification of Occupations) set and (d) Subgroups of ISIC (International Standard Industrial Classification of All Economic Activities) set are assigned.

While the national standard code number is sequential, the international standard code number has four hierarchical structures with Major groups, Sub-Major groups, Subgroups, and Unit groups. The numbers of the classified codes of (a), (b), (c) and (d) are nearly 200, 20, 400 and 60, respectively. Based on different requirement, a user can select any of the four code sets or a combination of them.

3.1 Methods

There are many kinds of methods in machine learning, such as decision trees or neural networks. SVMs are superior to the other methods in accuracy for many tasks including document classification

(Joachims, 1998; Sebastiani, 2002). In our work, we choose SVMs in the occupation and industry coding. We have to extend SVMs to a multi-classifier for the occupation coding because SVMs are a binary-classifier. For this purpose, we use the one-versus method (Kressel, 1999).

For the national standard codes we adopt the most effective combination method of SVMs and the rule-based method (Takahashi, 2005b), while for the international standard codes we adopt the methods of the SVMs as described in Section 3.1.1 because we have not developed a rule-based method for them.

3.1.1 Features

For the features in the methods of SVMs, we use words in response to "job task" (free answer), words in response to "industry" (free answer), and responses to "employment status" and "job title", which we call "basic features". The above term "words" means words as a part of speech that can be divided by morphological analysis (Kurohashi and Nagao, 1998). The words are translated into the feature numbers using a feature dictionary to which the set of a word and a number is automatically added in the case of newly appearing words in a free answer (Takahashi et al., 2005b).

In addition to basic features, the output of the rule-based method is used as a new feature to the national standard coding (Takahashi et al., 2005b). However, instead of the output of the rule-based method, the first-ranked predicted code is added to the international standard coding which has not the output of the rule-based method as a new feature. It was empirically shown that the method of SVMs adding the national standard codes as a feature was more effective for ISCO coding (Takahashi, 2008). Experiments in Takahashi (2008) also showed that the method using only the first-ranked predicted code was better than the method using the first-ranked and the second-ranked predicted codes together.

For ISCO coding, besides these features, "educational background" is also used as a new feature because educational background closely represents the concept of "skill level" as used in ISCO (ISCO, 1988; Tanabe and Aizawa, 2008).

3.1.2 Coding Processes

When the system assigns all four kinds of codes, the order of the process is SSM occupation coding, ISCO coding, SSM industry coding and ISIC coding (See Figure 1).

In SSM occupation coding, the system first conducts the rule-based method to adopt the method of SVMs. In the rule-based method, the system extracts a triplet of a verb firstly, the sub-categorized noun and the case of the noun from the responses to "job task" and "industry" (Takahashi, 2000). The case is a shallow case such as "wo" or "de", which are postpositional particles in Japanese. Second, the system searches for a rule that matches the generalized triplet after using both the verb thesaurus and the noun thesaurus. Third, for some occupation codes, the system checks other occupation variables such as "employment status", "job title" and "firm size". In the method of SVMs, the system trains the training dataset for SSM occupation codes with the features described in Section 3.1.1, and predicts codes from the first-ranked to the third-ranked.

For ISCO coding, the system adopts the method of SVMs, which trains the training dataset for ISCO with the features described in Section 3.1.1, and predicts the three codes of the first-ranked to the third-ranked. SSM industry coding is used in the same way as SSM occupation coding described in Section 3.1.1, and ISIC coding is used in the same way as ISCO coding also in Section 3.1.1.

If the system requests only ISCO coding, it conducts SSM occupation coding, except in the case of using the results of previous surveys, which is described in Section 3.3. For only ISIC coding, the same method is used as for only ISCO coding.

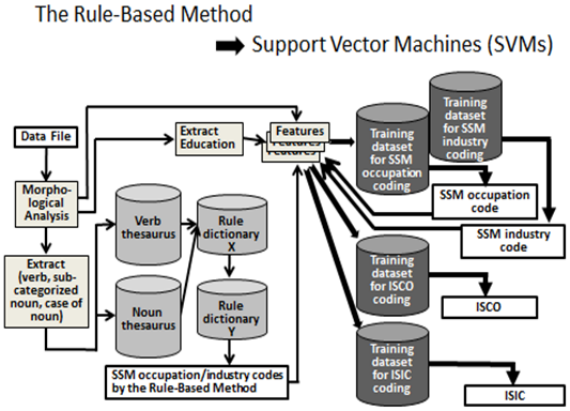

Figure 1: The process of the system.

3.2 Attaching a Three-Grade Confidence Level

Coders have asked us to supply a measure of confidence for the first-ranked candidate of their confidence decisions. For example, they can decide that there is no need for coding if the confidence

level is high. For this purpose, the system attaches a three-grade confidence level to the first-ranked predicted code by using classification scores.

A determining condition of the confidence level is as follows. In equations, Score1, Score2 represents a classification score for the first-ranked code, and a classification score for the second code, respectively.

Level A (high) which is the most useful to coders is determined on the condition that both (1) and (2) are satisfied. Level B (middle) is determined on the condition that both (1) and (3) are satisfied. Otherwise, Level C (low) is determined.

$$\text{Score1} > 0 \text{ and } \text{Score2} <= 0 \quad (1)$$

$$\text{Score1} - \text{Score2} > \text{Threshold} \quad (2)$$

$$\text{Score1} - \text{Score2} <= \text{Threshold} \quad (3)$$

In order to improve the prediction accuracy, we use not only a classification score for the first-ranked code, but also a score for the second-ranked code (Takahashi et al., 2008).

3.3 Using the National Standard Codes Determined by past Coding Tasks

According to the focus of international comparative studies (Jonsson, 2009; Iwai and Yasuda, 2011; Japanese Association of Social Research, 2011), international standard occupation and industry codes must be assigned to samples in previous surveys, too. For this purpose, the system assigns ISCO and ISIC by using SSM occupation codes and SSM industry codes already determined in past coding tasks, respectively.

In this case, a CSV data file contains an extra column for the determined code, which is added to data in a specified format. When the system obtains such a file, it assigns the international standard codes without the national standard coding. For example, if a CSV file contains SSM occupation codes, the system begins ISCO coding immediately.

3.4 System Performs

The CSV data file which a user sends to the Website of the SSJDA contains a "sample-ID", "education", "employment status", "job title", "job task" (free answer), "industry" (free answer) and "firm size" in this order. For example, a row of the CSV file is as follows: "sample-ID" is "1001", "education" is "9" (High school), "employment status" is "2" (Regular employee), "job title" is "1" (No managerial post), "job task" is "to arrange the delivery vehicles", "industry" is "load and unload of luggage" and "firm size" is "8" (From 500 to 999). If the CSV file does not contain clear information except "job task" ("industry") for occupation (industry) codes, the system can perform a task.

Figure 2 illustrates the GUI of the "an automatic coding system" (Japanese in the title bar of the screen), and those options in the middle of the screen GUI are "Occupation" (left side of the screen) and "Industry" (right side of the screen), respectively. Our system can identify the CSV data file automatically and filter the enable options of "Occupation/Industry" or "National/International" processing. Next, the system can run according to checked checkboxes by displaying the progress of the process.

Figure 2: The automatic coding system (a screen shot).

4 EXPERIMENTS

4.1 Experimental Settings

We conducted two kinds of experiments. In the first experiment (Experiment 1), we compared the performance of the system between four kinds of codes in terms of accuracy. In the second experiment (Experiment 2), we investigated the effectiveness of a three-grade confidence level in four kinds of codes.

Seven datasets are used in the experiments: JGSS-2000, JGSS-2001, JGSS-2002, JGSS-2003, JGSS-2005, JGSS-2006 and 2005SSM. Only the JGSS-2006 and 2005SSM datasets are assigned four kinds of codes in these datasets. The total number of the samples is 57412. We split a training dataset and a test dataset by assuming a real application, that is, a newer dataset is used as a test dataset and an older dataset is used as a training dataset without n-fold cross-validation (See Table 1). The total number of a training dataset and a test dataset for the national standard codes and the international standard codes is 39120 and 16089, respectively.

Table 1: A dataset for each code set.

Kind of codes	Training dataset	Test dataset
SSM occupation/ industry	JGSS-2000, -2001, -2002, -2003, -2005 (39120 samples)	JGSS-2006 (2203 samples), 2005SSM (16089 samples)
ISCO/ISIC	2005SSM	JGSS-2006

4.2 Experiment 1: Results and Discussion

Accuracy in this paper is defined as the number of correctly-classified samples divided by the number of all samples. We determined that the value of the goal of accuracies is more than 80%.

4.2.1 Direct Classification (the Current Method)

Table 2 shows accuracies up to the third-ranked of each kind of codes applied to the JGSS-2006 dataset and the 2005SSM dataset. In Table 2, codes with an asterisk* represent correctly classified national standard codes determined by past coding tasks. The same applies in Table 3, Table 5 and Table 6.

As for occupation codes which are more important for social surveys, we conducted the same experiments on the JGSS-2008 dataset (1357 samples) and the JGSS-2010 (2570 samples). Table 3 shows accuracies of them. The values in Table 3 are almost the same as those in Table 2.

Table 2: Accuracies of each kind of codes (JGSS-2006 and 2005 SSM).

Kind of codes	JGSS-2006	2005SSM
SSM occupation	78.8	80.6
SSM industry	90.8	91.6
ISCO	70.5	-
ISIC	80.1	-
ISCO*	74.8	-
ISIC*	86.2	-

Table 3: Accuracies of each kind of codes (JGSS-2008 and JGSS-2010).

Kind of codes	JGSS-2008	JGSS-2010
SSM occupation	78.9	78.3
ISCO	71.0	69.1
ISCO*	77.5	73.6

The accuracies in Table 2 are from 70% to 90%. While the highest of all accuracies is that of the SSM industry code set, the lowest is that of ISCO, which is lower than the value of the goal. The reason ISCO coding is the most difficult is because the number of its classified codes is the largest of all coding tasks. When comparing cases using correctly classified codes with cases using predicted codes, the use of correctly classified codes is higher as expected.

In the next Section, we conduct another experiment to improve the accuracy of ISCO.

4.2.2 Learning using a Hierarchy in the Classified Codes of ISCO

ISCO is constructed from nine Major groups which can be broken down to Sub-Major groups, Subgroups, Unit groups and "0" (armed forces). Therefore, we can conduct a method using a hierarchical structure which is two-step learning as follows. First, the system assigns a Major group to a sample and next, it assigns a Unit group to each Major group.

Table 4 shows the accuracies of the first-ranked in each Major group as applied to the JGSS-2006 dataset by comparing the method using a hierarchical structure (upper row) and the current method (lower row). The boldface number indicates the best accuracy of the two methods. Table 4 shows that the method using the hierarchical structure is almost as effective when a training dataset is larger, and even though the method using it is better than the current method, the difference of the two methods is small. The average of the accuracy of the method using the hierarchical structure is 65.5%, while that of the current method is 65.1%. Because of the smallness of each size of a training dataset, the method using the hierarchical structure cannot take effect.

Table 4: Accuracies in each Major group of two methods (ISCO).

Major group	1	2	3	4
Number of training dataset	311	1308	1308	2067
Method using a hierarchical structure	31.3	59.9	**56.1**	**77.6**
Current method	**35.8**	**61.3**	53.8	77.0

5	6	7	8	9
2458	867	1981	2135	988
77.9	**77.3**	**57.0**	**62.0**	53.2
77.3	75.2	55.8	**62.0**	**53.8**

The method using a hierarchical structure has the disadvantage that it always misleads a Unit group if

it mistakes a Major group at the first step. In fact, the accuracy for the first-ranked of the Major group is only 80.0% in this experiment. Furthermore, the method needs to conduct SVMs ten times, while the current method needs only one time. For these reasons, we do not change the learning method until the size of each training dataset of ISCO's Major groups is larger.

4.3 Experiment 2: Results and Discussion

The coverage in this section is defined as the number of assigned samples divided by the number of all samples. To decide the value of the Threshold in the formula of (2) or (3), we conduct an experiment which investigates both accuracy and coverage of samples attached at "level A" (the Y-axis) by changing the value of the Threshold (the X-axis) applied to the 2005SSM dataset (See Figure 3). We determine the value to 3 of the Threshold because it satisfies the condition that accuracy of more than 95% is desirable and coverage is as high as possible.

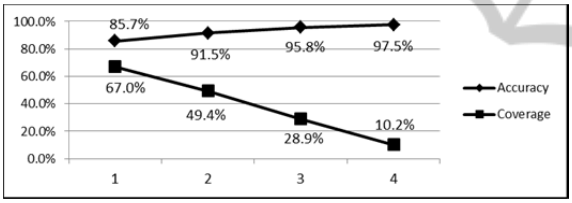

Figure 3: Accuracy and coverage in level A by changing the value of the Threshold.

Table 5 shows the accuracy and coverage (numbers in parentheses) applied to the JGSS-2006 dataset when the value of the Threshold is defined to 3. As for occupation codes, we conducted the same experiments on the JGSS-2008 dataset, the JGSS-2010 dataset and the 2005SSM dataset. Table 6 shows the accuracy and coverage (numbers in parentheses) of them. In Table 6, upper row, middle row and lower row show those of the JGSS-2008 dataset, the JGSS-2010 dataset and the 2005SSM dataset, respectively. The values in Table 6 are almost the same as those in Table 5.

Table 5 shows that in level A, accuracy is always higher than 94% and coverage is from 1% to 30%. While the accuracy of ISCO as a whole was 70.5% in Table 2, in level A it is 96.3%, which is satisfactory. Other kinds of codes are the same. Table 5 also shows that in level B, accuracies are always higher than 70% and these values are similar to those in Table 2. Accuracies in level C are always lower than those in Table 2. As a result, the confidence level can provide useful information for coders.

Table 5: Accuracy and coverage for each confidence level (JGSS-2006).

Kind of codes	Level A	Level B	Level C
SSM occupation	94.4(28.5)	70.8(47.2)	36.4(24.2)
SSM industry	97.3(31.5)	86.7(54.0)	43.7(14.5)
ISCO	96.2(4.8)	70.1(67.1)	27.6(28.1)
ISIC	94.1(0.8)	91.9(55.6)	57.4(43.6)
ISCO*	94.7(5.2)	75.9(64.7)	30.0(30.1)
ISIC*	100.0(0.7)	97.1(55.0)	67.1(44.3)

Table 6: Accuracy and coverage for each confidence level (JGSS-2008, JGSS-2010 and 2005SSM).

Kind of codes	Level A	Level B	Level C
SSM occupation	96.3(27.6) 94.0(30.4) 95.8(28.9)	71.4(46.6) 70.2(47.7) 71.9(47.8)	35.9(25.7) 31.7(21.9) 35.8(23.3)
ISCO	94.5(8.0) 90.7(5.4) -	62.6(50.4) 62.4(54.4) -	38.3(41.7) 40.1(40.2) -
ISCO*	96.6(8.5) 92.6(6.8) -	76.6(62.7) 73.9(65.4) -	32.4(28.7) 29.0(27.7) -

5 CONCLUSIONS

We have developed a new automatic coding system based on SVMs, which corresponds to each of the national/international occupation and industry standard code sets, for answers to open-ended questions. Second, we have attached a three-grade confidence level to each code predicted by the system by using multi-classification scores, and show that the confidence level is highly useful for coders. Third, we added a function of coding to the system for the international standard codes by using correct national standard codes already determined in past surveys. Finally, we have opened the system to the public through the SSJDA on the Web.

In future work, we need to improve the accuracy of ISCO so that it is higher than 80%. Next, we will add a new function to increase automatically training datasets for SVMs by using correctly-classified samples provided by future coding tasks. Finally, we would like to extend the system to other kinds of answers to open-ended questions.

ACKNOWLEDGEMENTS

The Japanese General Social Surveys JGSS were designed and carried out at the Institute of Regional Studies at Osaka University of Commerce in collaboration with the Institute of Social Science at Tokyo University. We received permission to use the 2005SSM dataset through the 2005SSM Survey Research Group. This research was partially supported by a MEXT Grant-in-Aid for Scientific Research (c) 25380640.

REFERENCES

1995SSM Survey Research Group, 1995. *SSM industry and occupation classification (the 1995 edition)*.

1995SSM Survey Research Group, 1996. *Codebook for 1995SSM survey*.

Hara, J., 1984. *Social surveys seminar,* University of Tokyo Press.

ISCO, 1988. http://www.ilo.org/public/english/bureau/stat/isco/isco88/ (accessed June 19, 2014)

ISIC, 1988. https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=2 (accessed June 19, 2014)

Iwai, N., Yasuda, T. (eds.), 2011. *Family values in East Asia: A comparison among Japan, South Korea, China, and Taiwan based on East Asian Social Survey 2006,* Nakanishiya Shuppan. Kyoto.

Japanese Association of social research, 2011. The special issue on difficulty and potential in cross-national survey research. *Advances in social research No.7,* Yuhikaku Publishing. Tokyo.

Joachims, T., 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning,* 137-142.

Jonsson, J. O. et al., 2009. Microclass mobility: Social reproduction in four countries. *American Journal of Sociology* 114(4), 977-1036.

Jung, Y., et al., 2008. A web-based automated system for industry and occupation coding. In *Proceedings of the Ninth International Conference on Web Information Systems Engineering,* 443-457.

Kressel, U., 1999. Pairwise classification and support vector machines. Scholkopf, B., et al. (eds.), *Advances in kernel methods support vector learning,* 255-268. The MIT Press.

Kurohashi, S., Nagao, M., 1998. *Japanese morphological analysis system JUMAN version 3.61,* Kyoto University.

NIOCCS, 2013. http://wwwn.cdc.gov/niosh-nioccs/ (accessed June 19, 2014)

Sebastiani, F., 2002. Machine learning automated text categorization. *ACM Computing Surveys* 34(1), 1–47.

Seiyama, K., 2004. *Social surveys,* Yuhikaku Publishing. Tokyo.

SOIC, 2000. http://www.cdc.gov/niosh/soic/default.html (accessed June 19, 2014)

SSJDA, 2013. http://ssjda.iss.u-tokyo.ac.jp/joint/autocode/ (accessed June 19, 2014)

Takahashi, K., 2000. A supporting system for coding of the answers from an open-ended question: An automatic coding system for SSM occupational data by case frame. *Sociological theory and methods* 15(1), 149-164.

Takahashi, K., 2002. An applying the automatic occupational/industrial coding system to JGSS-2000. Institute of Regional Studies, Osaka University of Commerce and Institute of Social Science, University of Tokyo (eds.), *JGSS Monographs No.1,* 171-184.

Takahashi, K., 2003. Applying the automatic occupational/industrial coding system to JGSS-2001. Institute of Regional Studies, Osaka University of Commerce, and Institute of Social Science, University of Tokyo (eds.), *JGSS Monographs No.2,* 179-191.

Takahashi, K., et al., 2005a. Classification of responses to open-ended questions with machine learning and hand-crafted rules: Automatic occupation coding methods. *Sociological theory and methods* 19(2), 177-196.

Takahashi, K., et al., 2005b. Automatic occupation coding with combination of machine learning and hand-crafted rules. *Lecture Notes in Artificial Intelligence* Vol.3518, 269-279. Springer. Heidelberg.

Takahashi, K., et al., 2005c. Applying the occupation coding supporting system for coders (NANACO) in JGSS-2003. Institute of Regional Studies, Osaka University of Commerce and Institute of Social Science, University of Tokyo (eds.), *JGSS Monographs No.4,* 225-241.

Takahashi, K., 2008. Automatic ISCO-88 coding with machine learning. *2005SSM survey series1 Problem in measurement and analysis in social surveys,* 47-68. 2005SSM Survey Research Group.

Takahashi, K. et al., 2008. Direct estimation of class-membership probabilities for multiclass classification using multiple scores. *Knowledge and Information System* 19(2), 185-210. Springer. London.

Tanabe, S., Aizawa, S., 2008. *University of Tokyo Institute of Social Science panel survey discussion paper series: An introduction manual and coding procedures for occupational and industrial coding system,* Institute of Social Science, University of Tokyo.

Todoroki, M., Sugino, I., 2013. *An introduction to social research methods,* Horitsu Bunkasha. Kyoto.

Vapnik, V., 1998. *Statistical learning theory,* Wiley. New York.