Computer-Assisted Historical Occupation Coding

Alex Gendlin
Center for Population Economics
1101 East 58th Street, Room 118
Chicago, Illinois 60637
773-753-0888
agendlin@cpe.uchicago.edu

Peter Viechnicki
Vredenburg
4831 Walden Lane
Lanham, Maryland 20706
301-306-2828
pviechnicki@vredenburg.com

ABSTRACT
The Center for Population Economics has recently implemented a recommender system for computer-assisted coding (CAC) of historical occupation descriptions. The system will be used to code occupation descriptions into standard occupation categories. We describe the task and the system, and discuss certain theoretical and practical issues which have arisen during its implementation, as well as their implications for Automated Text Categorization (ATC).

Categories and Subject Descriptors
I.5.4 [**Pattern Recognition**]: Applications – *Text Processing.*

I.5.5 [**Pattern Recognition**]: Implementation – *Interactive Systems.*

J.4 [**Social and Behavioral Sciences**]: *Economics.*

General Terms
Measurement, Design, Economics, Reliability, Human Factors.

Keywords
Text Categorization, Computer-Assisted Coding, *k*-Nearest Neighbor Classification, Machine Learning.

1. BACKGROUND

1.1 Data

As part of a National Institute on Aging program project entitled *Early Indicators of Later Aging, Disease, and Death,* the Center for Population Economics (CPE) between 1988 and 2000 collected a large textual database on the aging of the white veterans of the Union Army (UA). The data, drawn from a nationally-representative sample of 39,616 veterans, is intended to illuminate secular trends in, and determinants of, morbidity and mortality in the first cohort to have reached aged sixty-five in the twentieth century. The UA data consist of verbatim texts extracted from military records of the period 1835-1920,[3] and as such are replete with non-standard orthographies, circumlocutions, and data entry errors. The more than 25,000 unique occupational descriptions in the UA data range from single-word, or even single-letter tokens (e.g. 'FARMER', 'F'), to more colorful and complex descriptors (e.g. 'CUTTER AND LOGGER OF TIMBER').[1]

1.2 Code Sets

For use in economic and epidemiological modeling, raw occupational descriptions must be mapped onto discrete, mutually exclusive categories. This task is referred to as 'occupation coding.' Two standard category sets are currently used for this task. The first set, the Wilcox codes, grew out of studies of labor-force distribution in the antebellum economy,[8] and contains only nine categories. The second set, the 1950 U.S. Census Standard Occupation Codes,[5] is much richer in distinctions (n = 296), and has a two-level hierarchical organization of sub- and super-categories. Samples of both code sets are shown in Table 1.

Table 1. Sample Codes from Wilcox and 1950 Census Code Sets

Wilcox Code	Meaning	1950 Cen. Code	Meaning
1	Farmer	031	Dancers and Dancing Teachers
2	Professionals and Proprietors-I	032	Dentists
3	Professionals and Proprietors-II	033	Designers
4	Artisans	034	Dieticians and Nutritionists

As can be seen from the table, the 1950 Census codes are much narrower, and require extensive training to acquire, while the Wilcox codes are less specific, but more historically appropriate.

1.3 Task Constraints and Previous Processes

Two constraints on coding practice are imposed by the intended analytic purposes of the data: accuracy and recoverability/replicability. First, coding errors must be kept to an absolute minimum, so as to avoid introducing noise into the dataset. Second, all original data must be preserved, together with the rationale for each descriptor→code mapping; thus, should a

[1] The UA data are free and publicly available through the CPE: http://www.cpe.uchicago.edu.

later scholar choose to quibble with results based on coded occupation data, he or she could produce a different code mapping.

To code the existing UA sample given these overarching constraints, the CPE employed a laborious manual coding process between 1996 and 2001, whereby research assistants wrote transformation rules using a specialized regular expression language. The rewrite rules were then iteratively applied to the underlying data and edited, until an acceptable code mapping was achieved. For the Wilcox codes, an amorphous team of undergraduate research assistants wrote the transformation rules over a period of several years. For the 1950 Census codes, because of their complexity, it was necessary to employ a graduate economics student for more than one year to accomplish the mapping. This approach was costly, and (more crucially for the *Early Indicators* project) suffered from a lack of consistency over time and across the full dataset. For example, the two descriptors shown in Table 2 were coded differently, even though presumably the first mapping is the correct one:

Table 2. Inconsistent Coding Using 1950 Census Codes

Descriptor	Code
'ANYTHING SOLDIER CAN DO'	970: Laborers (n.e.c.)
'ANYTHING SOLDIER COULD DO'	595: Members of the armed services

This type of inconsistency is anathema to statistical modelers, as it introduces unexplained heterogeneity into the data, and reduces the significance of estimates derived from it.

2. CURRENT EFFORTS

The CPE is currently expanding its database to include records from a sample of more than 6,000 African American veterans of the Union Army. A review of the coding process determined that improvements over existing manual coding techniques were possible given recent advances in ATC.[6] A fully automated coding system was not selected, because even 90% accuracy was deemed too low, given the first constraint on data quality mentioned in Section 1.3. Instead, a computer-assisted coding system, known as 'Recommender,' has been developed in order to achieve the correct balance of the accuracy of human coding with the speed and consistency of computerized categorization.

2.1 CAC System Architecture

The architecture of the Recommender CAC system is shown in Figure 1. Recommender runs under Solaris 8 on a Sun Fire 280R Server with 2 1.015 GHz UltraSparc 3 Cu processors and 2 GB of RAM. The database management system is PostgreSQL, running under Red Hat Linux on a separate Pentium II workstation. The classifier training module and the classification engine are both implemented in Perl. The GUI is written in tcl/TK with custom C extensions, and runs on either Linux or Windows clients.[2]

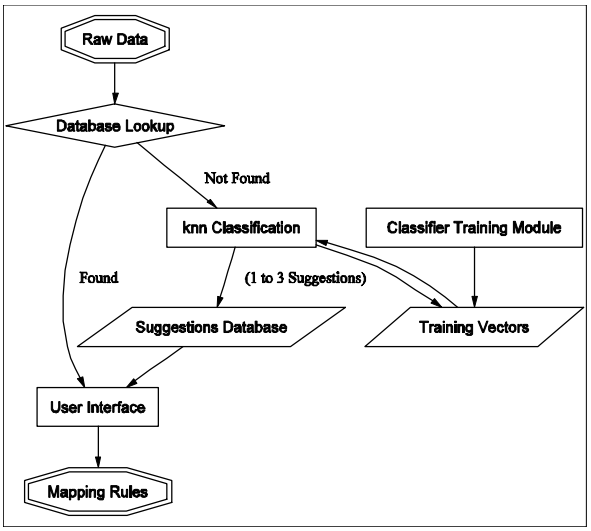

Figure 1. Recommender CAC System Architecture

The Recommender system works as follows: raw occupational descriptions are presented to the user on the left panel of the GUI; one to three suggestions for each of the two code sets are presented in the center and right-hand panel. The suggestions are generated by first querying the existing rule-base for exact matches for the uncoded occupation description; if one or more matches are found, associated codes are presented ordered by frequency. If no exact match is found, two *k*-nearest neighbor classifiers are used to select and suggest the most similar occupation descriptions in the training set. The coder either accepts or changes the recommendations of the CAC system. Even though both training and recommendation generation are comparatively fast (training time < 1 hour for 25,000 occupation descriptions; suggestion generation ≈ 5 seconds per description) both are done off-line to improve usability.

2.2 *k*-Nearest Neighbor Classification

The multiclass classifiers are trained on vector-space representations formed from features chosen according to the chi-squared criterion.[10] Features are chosen from the set of unique terms in the training data after stemming using the Porter stemmer. We currently select 80% of the stemmed terms as features, much higher than the percentage of features traditionally selected in ATC systems for longer texts (e.g. [10] finds 12.5% to be the optimal percentage). We find that selecting lower percentages of features leaves certain less common occupation categories with no characteristic features at all,[3] due to the comparative sparseness of the vector space formed from short texts.[4] Both training vectors and test vectors are formed using traditional *tf*idf* weights. A minor technical innovation employed here is to combine equivalent texts (description tokens) into single training vectors, preserving their count information. This practice

[2] All components of the system are freely available by special arrangement.

[3] This observation leads us to suspect that using unstemmed terms as features might lead to performance improvements, but we have yet to test this conjecture.

[4] Short document length has been shown to degrade performance in many types of information systems, e.g. [1].

has improved processing time by a factor of three, with no information loss.

While we have yet to undertake an exhaustive study of precision-recall tradeoffs as a function of *k,* our informal testing has led us to establish the current setting of $k = 20$, again differing from garden-variety *k*-nn classifiers. Our empirical results suggest that higher values of *k* sacrifice speed while adding little additional accuracy, but at values lower than 20, accuracy declines precipitously. The distance metric employed by the classifier is the standard cosine similarity between the vectors. Up to three suggestions are generated, using a simple voting procedure.[9] It is possible that more sophisticated techniques of selecting recommendations from amongst the nearest neighbors would result in higher accuracy (perhaps through the use of category-specific thresholds) however no such efforts have as yet been undertaken.

3. SYSTEM PERFORMANCE

The Recommender system's performance (before human correction) has been tested on both the Wilcox coding task and the 1950 Census coding task by reserving a subset of the coded training data as test data. 3,374 occupational description types were used as training data, corresponding to 39,910 tokens. Test results for both tasks are given in Table 3 below, with overall results for 1950 Census coding and Wilcox coding, and category-specific results given for the nine Wilcox occupation categories.

Table 3. Results of Automated Coding

Task/Category	Accuracy (correct/total)	Number Tokens
1950 Census Codes Overall	69%	2,177
Wilcox Codes Overall	77%	9,997
1. Agriculturalist	97%	1,299
2. Professionals and Proprietors-I	85%	852
3. Professionals and Proprietors-II	98%	1,323
4. Artisans	76%	2,131
5. Service, Semi-Skilled, and Operative	79%	1,089
6. Manual Labor	99%	2,222
7. Unidentifiable	49%	59
8. Unproductive	85%	747
9. Agricultural Labor	87%	274

Accuracy was calculated as a percentage of correct responses over total responses. For the Wilcox coding task, accuracy was calculated allowing only one guess per test vector: since there are only nine total categories in the Wilcox codes, allowing multiple predictions rapidly trivializes the performance evaluation. For calculating accuracy of the 1950 Census codes task, if any of the first three predictions matched the ground-truth category, the response was scored as correct.

4. IMPLICATIONS

One of the immediate findings from the results in Table 3 is that some occupational categories were more difficult for the system to acquire than others, such as Wilcox codes 4, 5, and 7. The same is true of the 1950 Census categories, though space limits prevent laying out all the relevant data. While the existence of category-level differences in classifier performance has long been noted in passing in the ATC literature, we believe it has important implications both for the current project in particular, as well as for ATC in general.

Human coders of the original occupational data also displayed category-level differences in consistency. For example, the Wilcox distinction between Professionals and Proprietors I and II, as well as the distinction between Artisans and Operatives, were both difficult for RAs to master consistently. We first wished to investigate whether the observed category-level differences were an artifact of classifier design, or whether the computer classifier displayed a similar pattern of confusions to the human pattern. Two Wilcox category confusion matrices were created, one from the Recommender system, and one from inconsistencies in the training data coded by hand. Multi-dimensional scaling was performed on the confusion matrices; the results of MDS in two dimensions are shown in Figure 2.

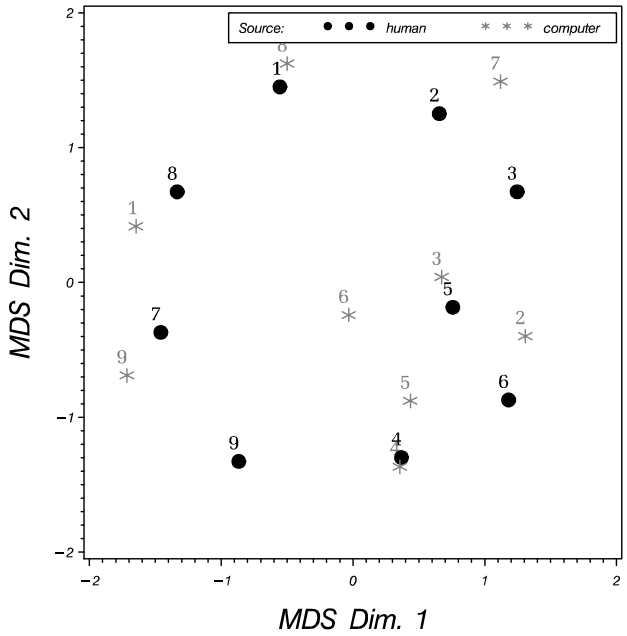

Figure 2: Multidimensional Scaling of Human and Computer Confusion Matrices, Wilcox Occupation Codes

Note: Computer dimension 2 has been inverted for comparison purposes.

In Figure 2, computer categories are shown by grey crosses, and human categories are shown with black dots, both labeled with the Wilcox code. The general relationship of the categories within the system appears to be approximately the same for the human coders as for the Recommender system. The only exception to this statement is for Code 7: Unidentifiable, but there is reason to believe that catchall or default categories are not well modeled by current machine-learning methods.[4] We conclude that the human and computer coders appear to have learned the

same general system of categories, as revealed by their similar confusion matrices.

It thus appears that the category-level differences in classifier performance are a property of the data and code sets themselves, and not of the Recommender system. Efforts are currently underway to identify determinants of those category-level differences. Linguistic factors, such as mean response length, relative category-specific vocabulary size, and category-specific word frequency, as well as non-linguistic factors such as mean category-specific income and socioeconomic status data, are currently being analyzed in order to measure their relative contributions to classifier performance.

The specific implication of this finding for occupation coding within the *Early Indicators* project is as follows. Coded occupation data is typically employed in statistical models as a proxy for some combination of wealth, income, socioeconomic status, prestige, and education. If category-level coding accuracy is correlated with any of these latent variables, then the coding process itself will introduce bias into the data. We are currently measuring the correlation between household wealth, the Duncan Socio-economic Index (SEI), and category-specific accuracy scores. Should either correlation prove significant, additional corrective steps will need to be put into the coding process to 'level the playing field' for all occupational categories. From a more general, human factors perspective, it will be desirable to design CAC systems in such a way as to help them focus the human coder's attention on potentially problematic cases, while still allowing them to speed through less ambiguous cases. Incorporating category-specific notions of suggestion confidence into the GUI is an obvious first step toward this goal.

The general implication of this finding for ATC is that evaluation metrics should be enriched to incorporate category-specific difficulty level. In other words, when scoring a CAC system for the Wilcox occupation codes, the system should get more points for each correct answer in category 7 than for each correct answer in category 6. Such information is not captured by any widely accepted ATC evaluation metric, except only very imprecisely by macro-averaged precision/recall. Without incorporating category-specific difficulty level, it will be impossible to define a stable, replicable measurement system for classifier performance.

5. RELATED WORK

The task of automating occupation coding can be seen as a subset of Automated Text Categorization, which is a well-described task (see e.g. [2], [6], [9]). Within ATC, occupation coding is most similar to Automated Survey Coding (ASC), which also operates on short texts for social-science research, and where accuracy (non-introduction of noise into the data) is paramount. The results presented here are an improvement over [7], though on a different dataset, and are comparable to results recently reported in [4]. Note that the Recommender system only accidentally shares a name with recommender systems used in e-commerce, but otherwise is quite different.

6. SUMMARY AND FUTURE RESEARCH

We have presented a description of the Recommender CAC system, which is currently being used at the CPE for historical occupation coding. We also have presented test results of the accuracy of the system's predictions. We discussed category-level differences in system performance, and the problems they raise for ATC scoring metrics as well as for the use of coded occupation data in statistical models which include socioeconomic status.

7. ACKNOWLEDGMENTS

We thank Grigoriy Abramov of the CPE for assistance in setting up hardware and software components necessary to this project, and Joseph Burton for managerial support.

8. REFERENCES

[1] Croft, W.B., S.M. Harding, K. Taghva, & J. Borsack. (1994). 'An Evaluation of Information Retrieval Accuracy with Simulated OCR Output.' *Symposium of Document Analysis and Information Retrieval.*

[2] Dumais, S., J. Platt, D. Heckermann, & M. Sahami. (1998). 'Inductive Learning Algorithms and Representations for Text Categorization.' http://research.microsoft.com/~sdumais/cikm98.pdf.

[3] Fogel, R. (1993). 'New Sources and New Techniques for the Study of Secular Trends in Nutritional Status, Health, Mortality, and the Process of Aging.' *Historical Methods* 26:5-43.

[4] Giorgetti, D., & F. Sebastiani. (2003). 'Automating Survey Coding by Multiclass Text Categorization Techniques.' *Journal of the American Society for Information Science and Technology,* (in press).

[5] Integrated Public-Use Microsample Documentation, variable OCC1950. http://www.ipums.umn.edu/usa/pwork/occ1950b.html.

[6] Lewis, D., & M. Ringuette. (1994). 'A Comparison of Two Learning Algorithms for Text Categorization.' *Third Annual Symposium on Document Analysis and Information Retrieval,* pp. 81-93.

[7] Viechnicki, P. (1998). 'A Performance Evaluation of Automatic Survey Classifiers.' In V. Honavar & G. Slutzki (eds.), *Proceedings of ICGI-98, 4th International Colloquium on Grammatical Inference.* Heidelberg, Germany: Springer Verlag.

[8] Wilcox, N. (1994). 'A Note on the Occupational Distribution of the Urban United States in 1860.' In Robert W. Fogel et. al. (eds.), *Without Consent or Contract: The Rise and Fall of American Slavery, vol. 2, Evidence and Methods.* New York: W. W. Norton and Co.

[9] Yang, Y., & X. Liu. (1999). 'A Re-examination of Text Categorization Methods.' *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR'99,* pp. 42-49.

[10] Yang, Y., & J. Pedersen. (1997). 'Feature Selection in Statistical Learning of Text Categorization.' *Fourth International Conference on Machine Learning,* pp. 412-420.