Mapping Text to Occupational Characteristics; Mapping Job Titles to SOC Codes; Mapping Job Titles to OCC Codes

Enghin Atalay, Phai Phongthiengtham, Sebastian Sotelo, Daniel Tannenbaum[1]

Abstract

In this document, we outline and justify the procedures we use first, to convert the text from job ads to different occupational characteristics, and second to map the job title to either an SOC code or an OCC code. For the first task of constructing mappings from job ad text to occupational characteristics, we rely on three complementary approaches: i) mappings based on our own judgment of the meaning of words and phrases, ii) mappings based on a continuous bag of words model which we have constructed, and iii) mappings used in previous papers. We describe these three approaches in order. Then, we discuss our methods to construct a mapping between job titles, SOC codes, and OCC codes.

Mappings between occupational characteristics and words or phrases

By this point, we have constructed a data set which contains job titles and the text within each job ad. From here, the next two tasks are i) to characterize the text within each job ad in terms of the skills, tasks, and other occupational elements described in the job ad. More precisely, our goal is to measure the prevalence of words and phrases related to the following characteristics:

1. O*NET work styles, skills, knowledge requirements, and activities;
2. Non-routine (interactive, analytic, and manual) and routine (cognitive, and manual) tasks;

[1] Atalay and Phongthiengtham: Department of Economics, University of Wisconsin-Madison. Sotelo: Department of Economics, University of Michigan-Ann Arbor. Tannenbaum: Department of Economics, University of Nebraska-Lincoln. We acknowledge financial support from the Washington Center for Equitable Growth.

3. Character, general computer skills customer service skills, financial skills, people management skills, problem solving skills, project management skills, social skills, and writing skills;

4. Personality trait-related words

5. Usage of different technologies (e.g., Microsoft Word, Cobol, Fortran); and

6. Experience and education requirements.

For job characteristics in (5) and (6), we search for regular expressions which specify these occupational characteristics. For example, when determining whether a particular job ad mentions Microsoft Word, we search for “msword,” or “microsoft” + “word.” Similarly, we search for phrases that contain the strings “1-2 ye*”, “1-3 ye*”, “1+ ye*”, or “one ye*” when determining whether employers seek employees with at least one year of relevant experience.

For characteristics in (1), (2), and (3) of the above list, we pursue a mix of approaches in constructing the list of words and phrases to search over.

Mappings used in previous papers

First, when possible, we appeal to lists used in previous papers. This is possible for the characteristics given in (2), (3), and (4) in the list above.

In constructing a set of words which relate to non-routine and routine task categories, we employ the categorization used by Spitz-Oener (2006). Quoting from our paper, we write:

> *Building on her [Spitz-Oener’s] mapping from survey question titles to task categories, we search for the following sets of words for each task category: 1) non-routine analytic: analyze, analyzing, design, designing, devising rule, evaluate, evaluating, interpreting rule*, plan, planning, research, researching, sketch, sketching; 2) non-routine interactive: advertise, advertising, advise, advising, buying, coordinate, coordinating, entertain, entertaining, lobby, lobbying, managing, negotiate, negotiating, organize, organizing, presentation, presentations, presenting, purchase, sell, selling, teaching; 3) non-routine manual: accommodate, accommodating, accommodation, renovate, renovating, repair, repairing, restore, restoring, serving; 4) routine cognitive: bookkeeping, calculate, calculating, correcting, corrections, measurement, measuring; 5) routine manual: control, controlling, equip, equipment, equipping, operate, operating.*

Second, in constructing a list of words which correspond to different skill-related words, we apply a set of definitions used in Deming and Kahn (2017). Quoting from the appendix of our paper, we write:

> *See Table 1 of Deming and Kahn (2017) for their list of words and their associated skills. Building on their definitions, we use the following rules 1) cognitive: analytical, cognitive, critical thinking, math, problem solving, research, statistics; 2) social: collaboration, communication, negotiation, presentation, social, teamwork; 3) character: detail oriented, meeting deadlines, multi-tasking, time management; 4) writing: writing; 5) customer service: client, customer, customer service, patient, sales; 6) project management: project management; 7) people management: leadership, mentoring, people management, staff, supervisory; 8) financial: accounting, budgeting, cost, finance, financial; 9) computer (general): computer, software, spreadsheets.*

Finally, we use John, Naumann, and Soto's (2008) categorization of words, as they map to the "Big 5" personality traits. In particular, John, Naumann, and Soto (2008) specify that

- "talkative," "assertive," "energetic," "outgoing," "outspoken," "dominant," "forceful," "enthusiastic," "show-off," "sociable," "spunky," "adventurous," "noisy", and "bossy" as representing an extroverted personality;

- "sympathetic," "kind," "appreciative," "affectionate," "soft-hearted," "warm," "generous," "trusting," "helpful," "forgiving," "good-natured," "friendly," "cooperative," "gentle," "unselfish," "praising," and "sensitive" as representing agreeableness;

- "organized," "thorough," "planful," "efficient," "reliable," "dependable," "conscientious," "precise," "deliberate," "painstaking," and "cautious" as representing conscientiousness;

- "stable," "calm," and "contented" as representing the opposite of a neurotic personality;

- and "imaginative," "intelligent," "original," "insightful," "curious," "sophisticated," "artistic," "clever," "inventive," "sharp-witted," "ingenious," "witty," "resourceful," and "wise" as representing openness to new experiences.

See Table 4.4 of their paper. We excluded a couple of the words which were on John, Naumann, and Soto (2008)'s lists, because of the special role they play in job ads: "responsible," because this word is overwhelmingly used to describe job responsibilities rather than potential employees' personality traits, and "active," because this word is used overwhelmingly to describe firms.

A definite and important advantage of employing categorizations developed elsewhere in the literature is that it we are utilizing already-established measurements to characterize occupations. On the other hand, there is no analogous existing lists of words for O*NET Elements (which is something we wish characterize, so that we may build on one of the leading and most comprehensive data sets of occupational characteristics). Moreover, it is possible that the lists developed elsewhere may, in our context, not fully represent the skill and task groups we wish to measure. For these two reasons, we apply two other procedures to identify words to search for in our job ad text.

Mappings based on a continuous bag of words model

We apply a statistical model, a *continuous bag of words model*, to extend our lists of words and phrases to search for. As we write in the body of the paper, with the aim of providing intuition about the goal of the continuous bag of words model, the model:

> *...is based on the idea that words are similar if they themselves appear (in text corpora) near similar words. For example, to the extent that "iv nurse," "icu nurse," and "rn coordinator" all tend to appear next to words like "patient," "care," or "blood"one would conclude that "rn" and "nurse" have similar meanings to one another.*

For additional background on continuous bag of words models, see the Appendix of our paper, Bengio et al. (2003), or Mikolov et al. (2013a, 2013b).

In our implementation, we construct our model by taking as our text corpora all of the text from job ads which appeared in our cleaned newspaper data, plus the raw text from job ads which were posted on-line in two months: January 2012 and January 2016.[2] In terms of parameterizing what it means for two words to be "near" one another, we choose a window length of 5 words. When searching for sets of words which are similar to one another, we restrict attention to words which appear at least five times in our text corpus.

With this vector representation, we can compute the (cosine) similarity between any word/phrase and any other word in our text corpus. This similarity measure equals the dot product of the vectors representing each word/phrase, divided by the magnitudes of these two vectors. The similarity measure varies between 0 (for totally dissimilar words) to 1. For

[2] EMSI (http://www.economicmodeling.com/about/) provided us these text, containing 4.2 million ads.

instance, the similarity between “researching” (which is one of the words mentioned by Spitz-Oener as a non-routine analytic task) and “investigating” equals 0.72. After “investigating,” “researching” is most similar to “reviewing,” “identifying,” and “resolving.” For each of the words belonging to each Spitz-Oener task category (and for each of the words belonging to each Deming and Kahn skill category), we construct a union of the list of words based on the words which have a cosine similarity greater than 0.55 and the list of words with the top 10 cosine similarity scores for the given Spitz-Oener/Deming and Kahn category. For each O*NET Element Title, we construct a list of similar words which have a cosine similarity greater than 0.45 along with the words with the top 10 cosine similarity scores for the given O*NET Element.[3]

Mappings based on our own judgment

Third, we construct our own mappings between words and phrases on the one hand and job characteristics on the other hand. For each O*NET Element, we begin by looking for words and phrases related to the occupational characteristic’s title or description we use our own judgment whether individual synonyms should be included or excluded. To give an example, we refer to the text of our paper, where we write:

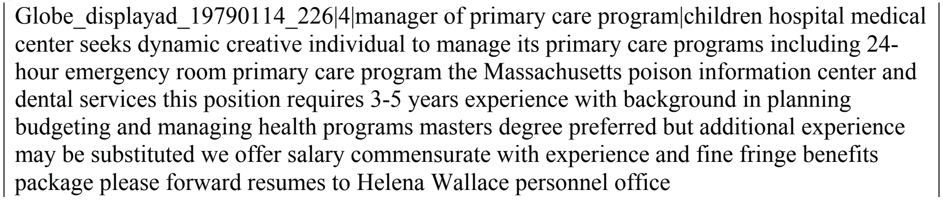
> *For instance, for the “Production and Processing” [O*NET] knowledge requirement, our list of synonymous words includes the original “production” and “processing,” and also “process,” “handle,” “produce,” “render,” and “assembly.” And, since the O*NET Description for “Production and Processing” states that the skill is associated with the “Knowledge of raw materials, production processes, quality control, costs, and other techniques for maximizing the effective manufacture and distribution of goods,” we also include “quality control,” “raw material,” “qc,” and “distribution” in our list of words and phrases to search for when measuring this knowledge requirement.*

Applying and combining these mappings: an example

The three sets of mappings are collected in the same website as the one hosting this document. For each mapping between a given occupational characteristic and a set of words and phrases, we count the number of occurrences in each job ad.

[3] We choose a higher threshold for the Spitz-Oener (2006) and Deming and Kahn (2017) categories as we are searching for words that are similar to *any* word in the lists that these previous authors have constructed.

Figure 1: Snippet of the January 14, 1979 Boston Globe, Display Ad #226



Globe_displayad_19790114_226|4|manager of primary care program|children hospital medical center seeks dynamic creative individual to manage its primary care programs including 24-hour emergency room primary care program the Massachusetts poison information center and dental services this position requires 3-5 years experience with background in planning budgeting and managing health programs masters degree preferred but additional experience may be substituted we offer salary commensurate with experience and fine fringe benefits package please forward resumes to Helena Wallace personnel office

We revisit the example from a previous document, a snippet of ads from the January 14, 1979 Boston Globe. We focus on ad #4 within this page of ads. This ad is for a “manager of primary care program,” the job title which we identified in a previous step. This ad contains

- two mentions each of the ” “Management of Personnel Resources,” “Operations Analysis,” and “Science” skills;
- three mentions of “Management of Financial Resources” skills
- two mentions of the “Personnel and Human Resources,” knowledge requirement;
- three mentions of the “Administration and Management” knowledge requirement;
- three mentions of “Organizing, Planning, and Prioritizing Work,” work activity;
- two mentions each of the “Monitoring Resources,” “Coordinating.” “Developing Objectives and Strategies,” and “Working with the Public” work activities
- two mentions each of the Spitz-Oener (2006) non-routine analytic and non-routine interactive task groups;
- two mentions of Deming and Kahn’s (2017) financial skill measure;
- and one mention each of several other occupational characteristics.

Where do these figures come from? Our own “judgment-based” mappings include “manage” and “managing” which relates to the “Administration and Management” knowledge requirement; “planning” which relates to the “Organizing, Planning, and Prioritizing Work” activity; and “budgeting” which relates to the “Monitoring Resources” work activity and the “Management of Financial Resources” skill.

Second, according to the mappings defined in Spitz-Oener (2006), we record one mention of non-routine analytic tasks based on the word “budgeting,” and one mention of non-routine interactive tasks based on the word “manage.” According to the Deming and Kahn (2017) definitions, “planning” also relates to Financial skill measure.

Third, our CBOW-model further indicates that the words “masters” and “degree” appear twice in representing “Science” skills. Also according to our CBOW-model, “planning” and “budgeting” relate to the “Operations Analysis” skill; “managing” and “services” to the “Management of Personnel Resources” skill; and “planning” to the “Management of Financial Resources” skill. Furthermore, our CBOW model relates the word “planning” to the “Administration and Management” knowledge requirement; the words “planning” and “programs” to the “Developing Objectives and Strategies” work activity and to the “Coordinating the Work and Activities of Others” activity; the words “managing” and “budgeting” to the “Organizing, Planning, and Prioritizing Work” work activity; the word “planning” to the “Monitoring Resources” work activity; and the words “preferred” and “experience” to the “Working with the Public” work activity.

Fourth, our CBOW-model indicates that the word “budgeting” refers to a non-routine analytic task, while “managing” refers to a non-routine interactive task. Our CBOW-model further indicates that “budgeting” relates to the Deming and Kahn (2017) financial skill measure.

Finally, from this ad we identify that a Masters degree and 3-5 years (which we record as 3 years) of experience is mentioned.

Note that, because different (existing in the literature) occupational measures measure similar work concepts, the same word will be recorded as representing different work measures: the word “managing” refers to the activity “Organizing, Planning, and Prioritizing Work” and the “Administration and Management” skill, and is classified as a non-routine interactive task.

Mappings between job titles and SOC codes[4]

By this point, we have constructed a data set which characterizes the occupational characteristics — using a manageable number of variables — of workers with different job titles. Our final step in our data set construction is to map individual job titles, e.g., manager of primary care program as in Figure 1, to SOC codes. This mapping serves two purposes. First, it will reduce the dimensionality of our data set in a useful way. There are several sets of observed job titles which pertain to the same occupation, but which are phrased in ever so slightly distinct ways. For example, the job titles “secretary to exec dir,” “secretary to division manager,” and “secretary to department manager ” appear in our newspaper text nine, six, and three times respectively. These job titles for all intents and purposes refer

[4] Much of the text of this section is taken directly from Appendix C.4 of our paper, *The Evolving U.S. Occupational Structure.*

to the same type of job. Recognizing this, it would be useful to map these similar types of jobs to one another. A second benefit of relating job titles to SOC codes is that the latter variable is measured in a number of previously existing data sets, for example the Decennial Census or the Current Population Survey. With a mapping between our job titles and SOC codes, one can merge our constructed database with these existing data sets.

Manually retrieving SOC codes for all of the job titles in our data set would be infeasible. There are, after all, more than 430 thousand unique job titles which are mentioned in at least two job ads, and nearly 99 thousand unique job titles which are mentioned in at least five job ads. So, we need some automated method to identify job titles' corresponding SOC codes. We retrieve SOC codes using our continuous bag of words model from the previous section. In particular, for each job title t in our newspaper data, we compute the similarity between t and all of the job titles, τ, which appear in O*NET's (version 22.1) either Sample of Reported Titles or Alternate Sample of Reported Titles. For each O*NET job title τ, we observe an SOC code. For the job title $t,$ we assign to t the SOC code of job title O*NET job title τ. We do this for any job title that appears at least twice in our newspaper data.

In a second step, we assign an SOC code of 999999 ("missing") if certain words or phrases appear — "associate," "career builder," "liberal employee benefit," "many employee benefit," or "personnel" — anywhere in the job title, or for certain exact titles: "boys," "boys boys," "men boys girls," "men boys girls women," "men boys men," "people," "professional," or "trainee." These words and phrases appear commonly in our newspaper ads and do not refer to the SOC code which our CBOW model indicates. "Associate" commonly appears the part of the name of the firms which are placing the ad. "Personnel" commonly refers to the personnel department to which the applicant should contact.

We also replace the SOC code for the job title "Assistant" from 399021 (the SOC code for "Personal Care Aides") to 436014 (the SOC code for "Secretaries and Administrative Assistants"). "Assistant" is the fifth most common job title, and judging by the text within the job ads refers to a secretarial occupation rather than one for a personal care worker. While we are hesitant to modify our job title to SOC mapping in an ad hoc fashion for any job title, mis-specifying this mapping for such a common title would have a noticeably deleterious impact on our dataset.

In a final step, we amend the output of the CBOW model for a few ambiguously defined job titles. These final amendments have no discernible impact on aggregate trends in task content, on role within-occupation shifts in accounting for aggregate task changes, or on the role of shifts in the demand for tasks in accounting for increased earnings inequality. First, for job titles which include "server" and which do not also include a food-service-related word — banquet, bartender, cashier, cocktail, cook, dining, food, or restaurant — we substitute an

SOC code beginning with 3530 with the SOC code for computer systems analysts (151121). Second, for job titles which contain the word "programmer," do not include the words "cnc" or "machine," we substitute SOC codes beginning with 5140 or 5141 with the SOC code for computer programmers (151131). Finally, for job titles which contain the word "assembler" and do not contain a word referring to manufacturing assembly work — words containing the strings electronic, electric, machin, mechanical, metal, and wire — we substitute SOC codes beginning with 5120 with the SOC code of computer programmers (151131). The amendments, which alter the SOC codes for approximately 0.2 percent of ads in our data set, are necessary our paper *New Technologies and the Labor Market*. Certain words refer both to a job title unrelated to new technologies as well as to new technologies. By linking the aforementioned job titles to SOCs that have no exposure to new technologies, we would be vastly overstating the rates at which food service staff or manufacturing production workers adopt new ICT software. On the other hand, since these ads represent a small portion of the ads referring to computer programmer occupations, lumping the ambiguous job titles with the computer programmer SOC codes will only have a minor effect on the assessed technology adoption rates for computer programmers.

Mappings between job titles and OCC (2000-2004) codes

If one wanted to link our newspaper-based occupational measures to data from the Decennial Census, and had access only to a mapping between job titles to SOC codes, then it would be necessary to construct an additional mapping between SOC codes and OCC codes. This is somewhat inconvenient, as the relationship between SOC codes and OCC codes is many to many.

For this reason, we construct a second mapping between job titles and OCC codes. In the previous section, we drew on a sample job title from O*NET's (version 22.1): the union of the Sample of Reported Titles or Alternate Sample of Reported Titles. Instead, here, we construct a list of job titles included in the Census 2000 Occupation Index.[5]

As with our job title to SOC mapping, for each job title t in our newspaper data, we compute the similarity between t and all of the job titles τ which appear in the Census 2000 Occupation Index. For the job title $t,$ we now assign to t the OCC code of job title OCC job title τ. We do this for any job title that appears at least twice in our newspaper data.

Also as with our job title to SOC mapping, we make two amendments. First, we assign

[5] See https://www.census.gov/topics/employment/industry-occupation/guidance/indexes.html .

an OCC code of 999 (“missing”) if certain words or phrases appear. Again, these are: “associate,” “career builder,” “liberal employee benefit,” “many employee benefit,” or “personnel” — anywhere in the job title, or for certain exact titles: “boys,” “boys boys,” “men boys girls,” “men boys girls women,” “men boys men,” “people,” “professional,” or “trainee.” We also replace the OCC code for the job title “Assistant” to 570 (the OCC code for “Secretaries and Administrative Assistants”), and the OCC code for the job title “Salesperson” to 476 (the OCC codes for “Retail Salespersons”).

We also amend the output of the CBOW model for the same ambiguously defined job titles. First, for job titles which contain the word “programmer,” do not include the words “cnc,” “machin*,” “numerical control,” or “machine,” we substitute SOC codes beginning with 5140 or 5141 with the OCC code for computer programmers (101). Also, for job titles which contain the word “assembler” and do not contain a word referring to manufacturing assembly work — words containing the strings electronic, electric, machin, mechanical, metal, and wire — we substitute OCC codes from 700 to 900 with the OCC code of computer programmers (101).

We finish with a warning: While we have checked the accuracy of our job title to SOC mapping by comparing SOC occupations’ O*NET measures to the analogous measures based on our newspaper data, it is impossible to perform the same comparison with the OCC-based occupation classification. Moreover, we have reason to believe that the job title to SOC mapping may be more accurate than the job title to OCC mapping. In particular, there are substantially fewer job titles within the Census 2000 Occupation Index (giving mappings between job titles and OCC codes) than there are in the union of the BLS Sample of Reported Titles or Alternate Sample of Reported Titles (giving mappings between job titles and SOC codes): 24 thousand versus 43 thousand. Since there are fewer job titles (for which we observe occupation codes) to match to with the OCC classification scheme, it is possible that this will lead to a less accurate final mapping between job titles that we observe in our newspaper data to occupation codes.

References

[1] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. “A Neural Probabilistic Language Model.” *Journal of Machine Learning Research*, 3: 1137-1155.

[2] Deming David, and Lisa B. Kahn. 2017. “Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals.” *Journal of Labor Economics*, forthcoming.

[3] John, Oliver P., Laura P. Naumann, and Christopher J. Soto. 2008. “Paradigm Shift to the Integrative Big Five Trait Taxonomy.” *Handbook of Personality: Theory and Research*, 3: 114-158.

[4] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. “Efficient Estimation of Word Representations in Vector Space.” Mimeo.

[5] Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. “Distributed Representations of Words and Phrases and Their Compositionality.” In *Advances in Neural Information Processing Systems*, 3111-3119.

[6] Spitz-Oener, Alexandra, 2006. “Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure.” *Journal of Labor Economics*, 24(2): 235-270.