JobLex: A Lexico-Semantic Knowledgebase of Occupational Information Descriptors

Koustuv Saha, Manikanta D. Reddy, and Munmun De Choudhury

Georgia Institute of Technology, Atlanta, United States
(koustuv.saha, mani, munmund)@gatech.edu

Motivation, Objective, and Summary. Technological advancements in several work sectors have influenced evolution of the landscape of work at an unprecedented speed, leading to the demand of continuous skill development [1,8]. In turn, this interests a number of stakeholders spanning across academia and industry in a number of disciplines including labor economics, who leverage large-scale data available from a variety of offline and online sources (e.g., resumes, job portals, professional social networking such as LinkedIn, search engine, job databases, etc.) [9,11,12]. On these data streams, describing job aspects and skills vary extensively, confounded by factors such as self-presentation, subjective perspectives on soft and hard skills, audience, and intrinsic traits such as personality and mindset [2,4,7,15,17]. Such data analyses require a taxonomy of keywords that are associated with skills per job description or type. However, most databases are only limited — they do not capture variants, typos, abbreviations, or internet slangs that are used on social media or in informal settings [6]. To facilitate research in this space, our work builds on a well-validated dictionary of occupational descriptors (O*Net) to propose a method, and correspondingly a knowledgebase, JobLex of occupational descriptors that can be used in computational social science and organizational studies [13]. We publish both our script and an example lexicon (for Twitter) for purposes of research and practical application.

JobLex. We obtain a database of occupational descriptors, Occupational Information Network (O*Net). O*Net (*onetonline.org*) is developed under the sponsorship of the U.S. Department of Labor/Employment and Training Administration, and has extensively been used in research [3,5,16]. It enlists and describes eight primary occupational categories expanded further into 248 leaf occupational-categories. The hand-curated occupational descriptors allow us to represent occupational descriptors in a theoretically-grounded fashion. To capture the linguistic and semantic context of these descriptors, we use word embeddings. In particular, we expand them into clusters of keywords on the basis of pre-trained word embeddings (GloVe) [10] in the lexico-semantic latent space of word-vector dimensions [14]. In our specific case, we choose 30 keywords per cluster (ranked on cosine similarity), and use the n-dimensional (n=200) word-vectors trained on word-word co-occurrences in a Twitter corpus of 6B tokens [10] (see Table 1 for example keywords in eight broad occupational descriptors). We qualitatively inspect JobLex to observe that its keywords are theoretically and intuitively associated with the categories that they belong to – for example, *understanding, feelings, person* occur with high similarity with *Concern for Others*, and *responsibilities, challenges, willingness* occur with high similarity with *Work Styles: Initiatives* [3]. For research and practical purposes, we publish the script and lexicon of JobLex at *github.com/joblex/joblex*.

Table 1: Job aspect types with their descriptions as obtained from O*Net.

Job Aspect	Example Keywords
Interests	people, think, working, learning, teaching, business, involved, reason, helping
Knowledge	development, technology, teaching, training, education, information, improve
Skills	people, learning, lesson, education, bridging, differences, behavior, intentions
Wk. Activities	spending, teaching, conflicts, resolving, disputes, performance, relationships
Wk. Context	people, competitive, require, group, offer, think, person, experience, schedule
Wk. Styles	working, understand, right, difficult, responsibilities, positive, improving, effort
Wk. Values	business, ability, allow, decisions, potential, development, leadership, honest

References

1. Stevie Chancellor and Scott Counts. Measuring employment demand using internet search data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 122. ACM, 2018.
2. Erving Goffman et al. The presentation of self in everyday life. 1959.
3. Maarten Goos, Alan Manning, and Anna Salomons. Explaining job polarization: the roles of technology, offshoring and institutions. *Offshoring and Institutions (December 1, 2011)*, 2011.
4. Jamie Guillory, Jason Spiegel, Molly Drislane, Benjamin Weiss, Walter Donner, and Jeffrey Hancock. Upset now?: emotion contagion in distributed groups. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 745–748. ACM, 2011.
5. Heike Heidemeier and Klaus Moser. Self–other agreement in job performance ratings: A meta-analytic test of a process model. *Journal of Applied Psychology*, 2009.
6. Louis Hickman, Koustuv Saha, Munmun De Choudhury, and Louis Tay. Automated tracking of components of job satisfaction via text mining of twitter data. In *ML Symposium, SIOP*, 2019.
7. Bernie Hogan. The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386, 2010.
8. Shagun Jhaver, Justin Cranshaw, and Shagun Counts. Measuring professional skill development in u.s. cities using internet search queries. In *ICWSM*, 2019.
9. Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2019.
10. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
11. Koustuv Saha, Ayse Elvan Bayraktaraglu, Andrew Campbell, Nitesh V Chawla, et al. Social media as a passive sensor in longitudinal studies of human behavior and wellbeing. In *CHI Ext. Abstracts*, 2019.
12. Koustuv Saha, Manikanta D Reddy, Vedant Das Swain, Julie M Gregg, Ted Grover, Suwen Lin, Gonzalo J Martinez, Stephen M Mattingly, et al. Imputing Missing Social Media Data Stream in Multisensor Studies of Human Behavior. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, 2019.
13. Koustuv Saha, Manikanta D Reddy, Stephen M Mattingly, Edward Moskal, Anusha Sirigiri, and Munmun De Choudhury. Libra: On linkedin based role ambiguity and its relationship with wellbeing and job performance. *PACM HCI*, (CSCW), 2019.
14. Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *ICWSM*, 2018.
15. Meredith M Skeels and Jonathan Grudin. When social networks cross boundaries: a case study of workplace use of facebook and linkedin. In *Proceedings of the ACM 2009 international conference on Supporting group work*, 2009.
16. Prasanna Tambe and Lorin M Hitt. Now it's personal: Offshoring and the shifting skill composition of the us information technology workforce. *Management Science*, 58(4):678–695, 2012.
17. Hui Zhang, Munmun De Choudhury, and Jonathan Grudin. Creepy but inevitable?: the evolution of social networking. In *Proc. CSCW*, 2014.