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 [1,7]. 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.) [8,10,11]. 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 [2,4,6,13,15]. 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. 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. 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 [3,5,14]. 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) [9] in the lexico-semantic latent space of word-vector dimensions [12]. 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 [9] (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.

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