Retrieving Skills from Job Description: A Language Model Based Extreme Multi-label Classification Framework

Akshay Bhola, Kishaloy Halder, Animesh Prasad and Min-Yen Kan
December 8, 2020
70% of the global workforce is made up of passive talent who aren’t actively job searching, and the remaining 30% are active job seekers.

87% of active and passive candidates are open to new job opportunities.

Online job boards (60%) are one of the most popular channel amongst job seekers to find desired jobs.

Reference: business.linkedin.com
Image Reference: wsj.com
Online job portals

• These portals receive thousands of applications, among which fewer than 10% candidates have appropriate skills

• Crucial task is matching suitable candidates with jobs

• **Popular approach**: matching the candidate’s associated set of skills by the job descriptions (JDs)

---

Issue arises due to the inconsistencies in job descriptions
Sample Job Description

Job description

Looking for a dynamic individual keen to join a growing organization in the fast-paced and expanding Supply Chain Software Industry. As Singapore goes through the digital revolution, the individual will play a paramount role in developing and upgrading our SaaS tools and valued added services for our clients.

Responsibilities:
1. Work with Architect, Framework Designer to develop the next generation of Cloud base Micro Services Suite
2. Design and implement Micro Service modules primary for Supply Chain systems. Design must be flexible and scalable for any future expansion or upgrade.

...

Bhola et al.

Required skills

1. C#
2. Agile Methodologies
3. AJAX
4. ASP.NET
5. .NET
6. C++
7. HTML
8. Java

Incomplete skill set (individually)
Note: The job descriptions might connotate the skill requirement for job which aren’t explicitly mentioned in the textual descriptions

- Simple string-matching algorithm can’t extract implicit skills
- The task requires more complex algorithms that can infer implicit skills from the job descriptions

Distribution of the count of skills from mycareersfuture dataset with their corresponding implicit occurrence
Proposed Method

We approach this task as a multi-label classification problem

Job description
Requirements performing end end software development cycle coding using Java j2ee spring framework Oracle pl SQL Multithreading angular js hibernate rest soap api oracle databases shell scripting degree Information Technology Engineering background minimum 5 9 years experience information technology software development must proven experience ……

Predicted skills
1. SQL
2. Java
3. XML
4. JavaScript
5. Software Development
6. Web Services

BERT-XMLC

BERT Encoder

Probability vector

σ
mycareersfuture Dataset

• No large-scale job description dataset was available
• We have collected data from Singaporean government website, mycareersfuture.sg of over 20,000 richly structured job posts
• Dataset contains 16 different fields of information

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of job posts</td>
<td>20,298</td>
</tr>
<tr>
<td>Number of distinct skills</td>
<td>2,548</td>
</tr>
<tr>
<td>Number of skills with 20 or more mentions</td>
<td>1,209</td>
</tr>
<tr>
<td>Average skill tags per job post</td>
<td>19.98</td>
</tr>
<tr>
<td>Average token count per job post</td>
<td>162.27</td>
</tr>
</tbody>
</table>

Table 1: mycareersfuture dataset statistics
Performance Comparison

- BERT-XMLC model outperforms all other state-of-the-art models on Recall@M metric for different $M$
- Similar trend is observed for MRR & $nDGC@M$ metrics for different $M$, with BERT-XMLC outperforming all other models

Recall Comparison (Recall@M for various $M$)
We proposed *Correlation Aware Bootstrapping (CAB)* process to further enhance the performance of ML models by taking into the structured representation of skills and their co-occurrences.

**Semantic Representation of Skills:**

\[ D_{sem} = \{ < s, encode_{OH}(s) > \mid s \in S \} \]

**Co-occurrence based Correlation of Skills:**

\[ D_{corr} = \{ < concatenate(s \in S_k), label(k) > \mid k \in \{1, \ldots, M\} \} \]
Table 2: Performance comparison with added CAB implemented models

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR</th>
<th>Recall</th>
<th>nDGC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>@5</td>
<td>@10</td>
</tr>
<tr>
<td>CNN-Kim</td>
<td>0.8195</td>
<td>16.38</td>
<td>28.21</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8417</td>
<td>16.83</td>
<td>29.27</td>
</tr>
<tr>
<td>BiLSTM</td>
<td>0.8565</td>
<td>17.51</td>
<td>30.32</td>
</tr>
<tr>
<td>BiGRU</td>
<td>0.8716</td>
<td>17.73</td>
<td>30.80</td>
</tr>
<tr>
<td>BiGRU w/ CSA</td>
<td>0.8840</td>
<td>18.34</td>
<td>31.71</td>
</tr>
<tr>
<td>BERT-XMLC</td>
<td>0.9019</td>
<td>19.60</td>
<td>33.64</td>
</tr>
<tr>
<td>BIGRU w/ CSA + CAB</td>
<td>0.8995</td>
<td>18.93</td>
<td>32.72</td>
</tr>
<tr>
<td>BERT-XMLC + CAB</td>
<td>0.9049</td>
<td>21.67</td>
<td>35.93</td>
</tr>
</tbody>
</table>
Micro-evaluation metrics

Explicit Inference Measure (EIM): Instance-based measure of explicit skills predicted by the model, compared against gold-standard explicit skills mentioned for a JD

\[
EIM = \frac{\text{# explicit skills predicted by model}}{\text{# explicit skills mentioned in JD}}
\]

Note: EIM metric can have a value more than 1
Micro-evaluation metrics

**Relative Implicit Inference Measure (RIIM):** Recall-based measure of implicit skills predicted by the model, relative to the entire set of implicit skills

\[
RIIM = \frac{\# \text{ implicit skills predicted by model}}{\text{total implicit skills in JD}}
\]

**Relative Explicit Inference Measure (REIM):** Recall-based measure of explicit skills predicted by the model compared to the entire set of explicit skills

\[
REIM = \frac{\# \text{ explicit skills predicted by the model}}{\text{total explicit skills in JD}}
\]
Performance Comparison

Performance comparison of different models over EIM, RIIM, REIM metrics
Conclusion

• Collected a large-scale job description dataset – mycareersfuture*
• Proposed a novel transformer based model to handle the XMLC task
• Proposed model outperforms all other state-of-the-art XMLC models on the mycareersfuture dataset
• Proposed Collaboration Aware Bootstrapping technique to boost the performance of ML models in multi-label classification task

* Repo: https://github.com/WING-NUS/JD2Skills-BERT-XMLC
Thank you for your attention

Akshay Bhola
Email ID: akbhola.bhola@gmail.com
For any query, feel free to contact me.