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

NUS National University of Singapore
School of Computing

COLING 2020

WING
Web Information Retrieval
Natural Language Processing Group

Job Search in the Digital-era

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.

….

Incomplete skill set (individually)

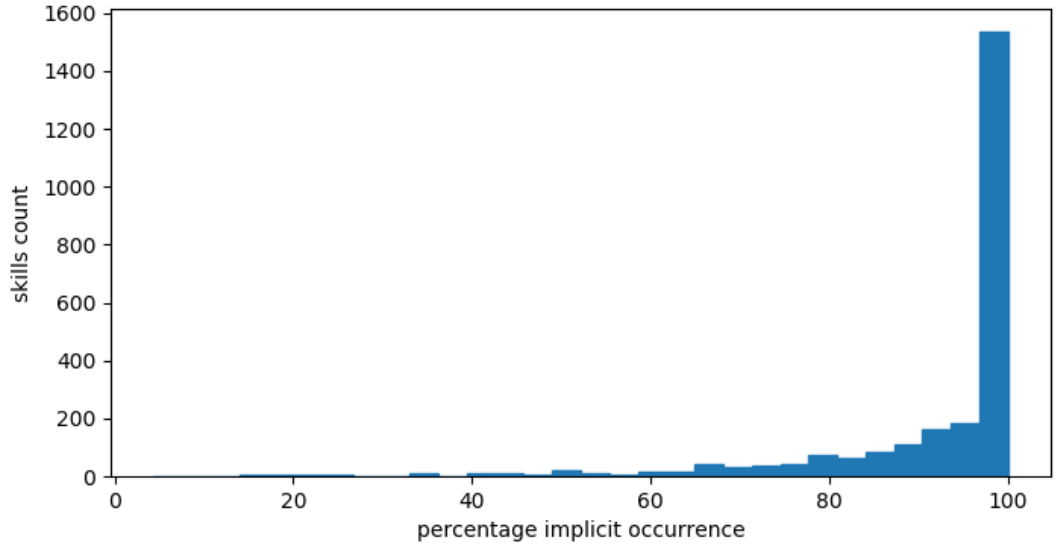
Required skills

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

…

Complexity of the task

Note: The job descriptions might connotate the skill requirement for job which aren't explicitly mentioned in the textual descriptions

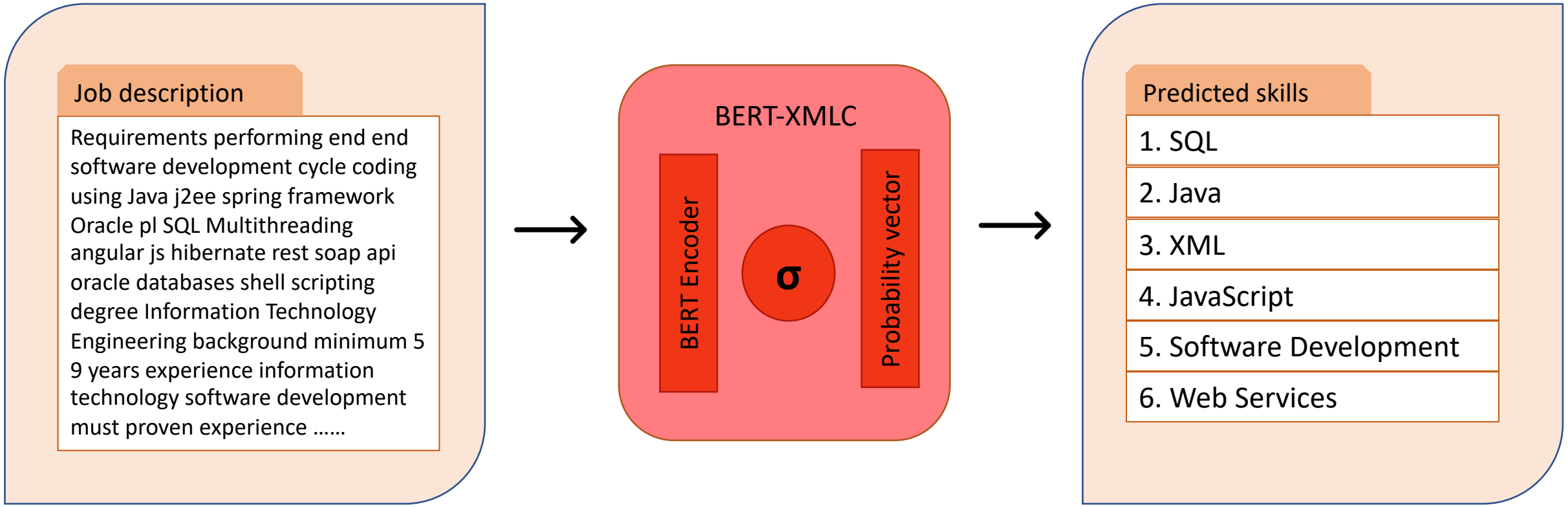


Distribution of the count of skills from mycareersfuture dataset with their corresponding implicit occurrence

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

Proposed Method

We approach this task as a multi-label classification problem



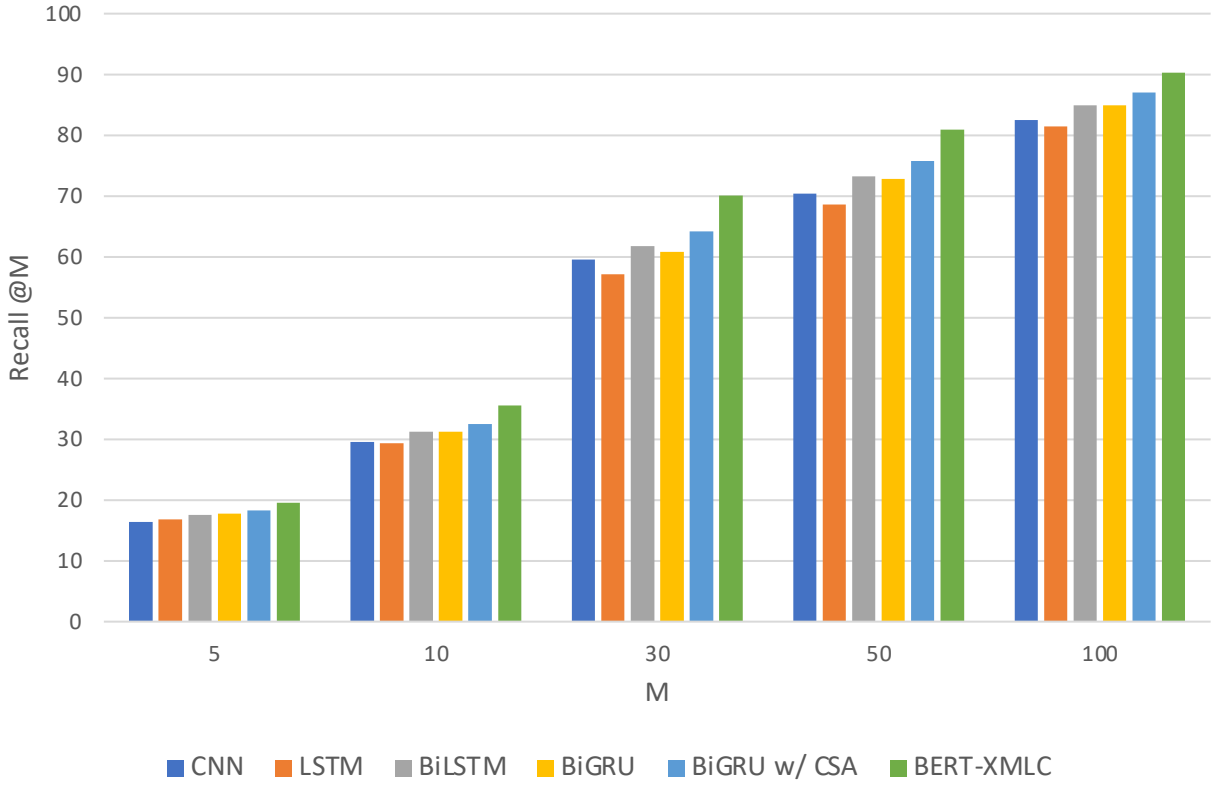
mycareersfuture Dataset

- 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

Number of job posts	20,298
Number of distinct skills	2,548
Number of skills with 20 or more mentions	1,209
Average skill tags per job post	19.98
Average token count per job post	162.27

Table 1: mycareersfuture dataset statistics

Performance Comparison



Recall Comparison (*Recall@M* for various *M*)

- 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

Correlation Aware Bootstrapping

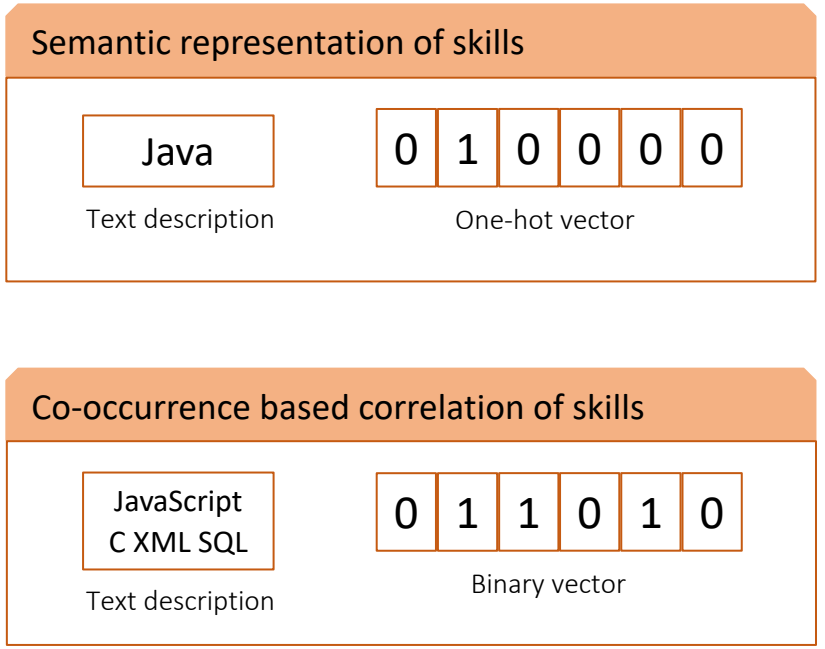
We proposed *Correlation Aware Bootstrapping (CAB) process* to further enhance the performance of ML models by taking into the structured representation of skills and their co-occurrences

Semantic Representation of Skills:

$$D_{sem} = \{< s, encode_{OH}(s) > \forall\ s \in S\}$$

Co-occurrence based Correlation of Skills:

$$D_{corr} = \{< concatenate(s \in S_k), label(k) > \forall\ k \in \{1, .., M\}\}$$



Added boost with CAB

Table 2: Performance comparison with added CAB implemented models

Model	MRR	Recall | nDGC									
		@5		@10		@30		@50		@100	
CNN-Kim	0.8195	16.38	28.21	29.59	40.23	59.60	60.60	70.40	66.37	82.47	71.96
LSTM	0.8417	16.83	29.27	29.35	40.68	57.09	59.43	68.65	65.61	81.46	71.53
BiLSTM	0.8565	17.51	30.32	31.19	42.77	61.77	63.50	73.25	69.64	84.89	75.04
BiGRU	0.8716	17.73	30.80	31.27	43.11	60.84	63.09	72.80	69.49	84.88	75.09
BiGRU w/ CSA	0.8840	18.34	31.71	32.52	44.62	64.19	66.04	75.75	72.23	87.02	77.46
BERT-XMLC	0.9019	19.60	33.64	35.58	48.18	70.10	71.66	80.91	77.45	90.26	81.79
BIGRU w/ CSA + CAB	0.8995	18.93	32.72	33.84	46.28	66.36	68.33	77.66	74.38	88.28	79.30
BERT-XMLC + CAB	**0.9049**	**21.67**	**35.93**	**40.49**	**52.84**	**79.59**	**79.32**	**86.60**	**82.96**	**92.24**	**85.41**

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

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

Note: EIM metric can have a value more than 1

Micro-evaluation metrics

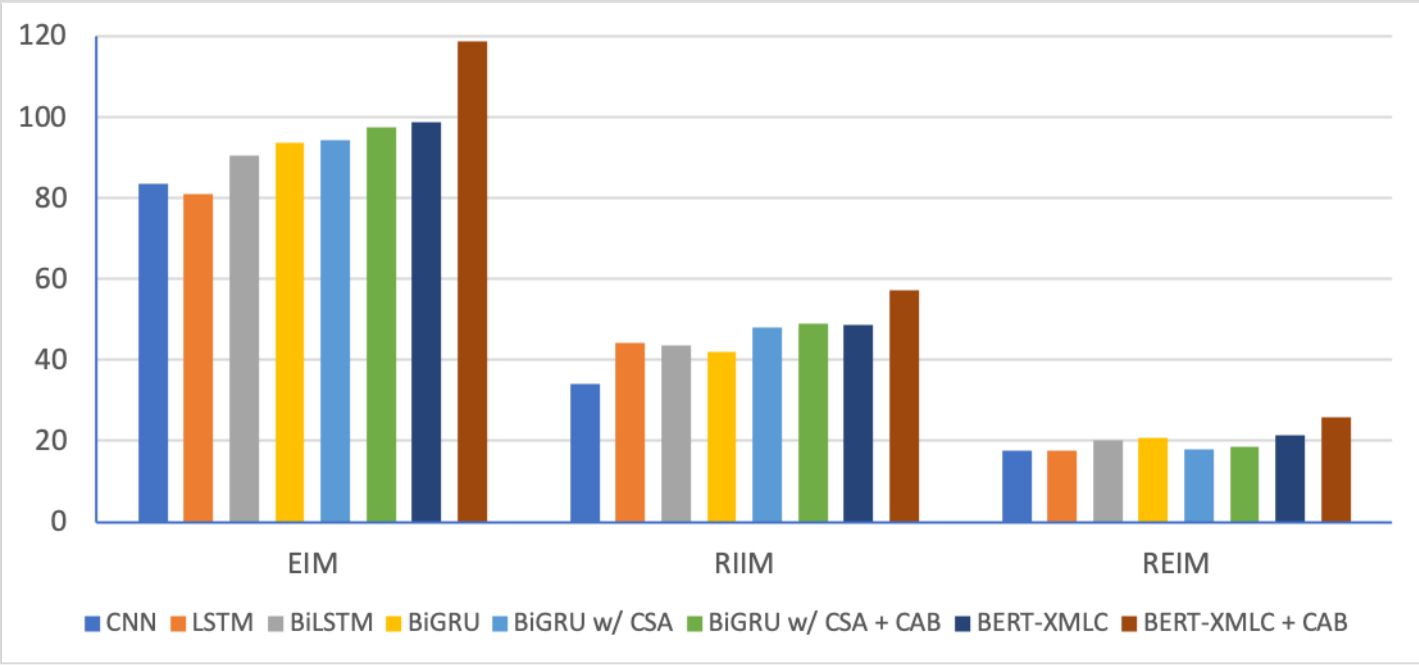
Relative Implicit Inference Measure (RIIM): Recall-based measure of implicit skills predicted by the model, relative to the entire set of implicit skills

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

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

$$\text{REIM} = \frac{\#\ explicit\ skills\ predicted\ by\ the\ model}{total\ explicit\ skills\ in\ JD}$$

Performance Comparison



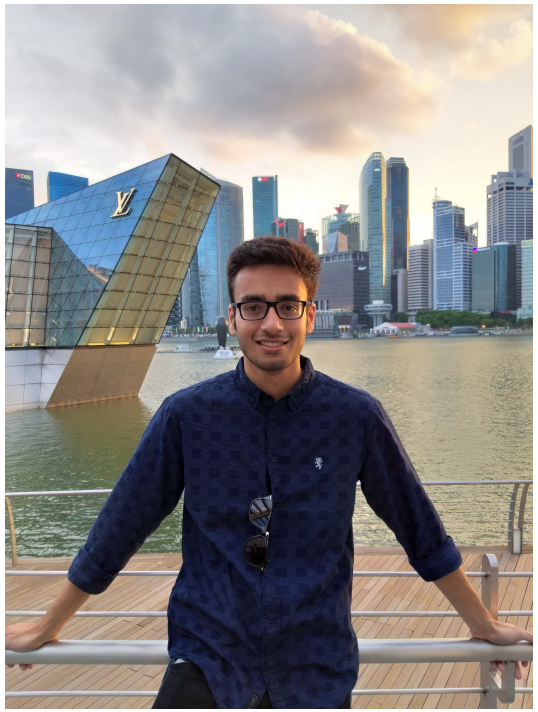
Performance comparison of different models over EIM, RIIM, REIM metrics

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.