JPLink: On Linking Jobs to Vocational Interest Types

Amila Silva, Pei-Chi Lo, and Ee-Peng Lim
amila.silva@student.unimelb.edu.au,
{pclo.2017@phids.,eplim@}smu.edu.sg

Living Analytics Research Centre
Singapore Management University

Pacific-Asia Conference on Knowledge Discovery and Data Mining
Singapore
May 11-14, 2020

Motivation

- Linking job seekers with relevant jobs requires matching based on:
  - job criteria (e.g., skills, abilities, knowledge, etc.)
  - **personality types**
- Holland Code (i.e., RIASEC)[1] is generally used to characterize personality types of jobs and applicants

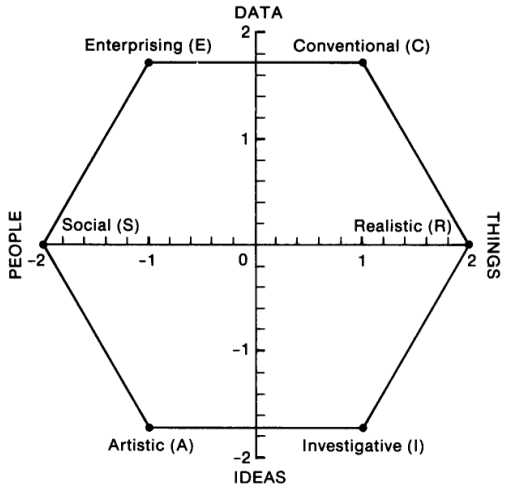


Figure: The Hexagonal Model of Holland's Vocational Interest Types (source:[2])

[1] John L Holland. *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources, 1997.

[2] Dale J. Prediger. "Mapping Occupations and Interests: A Graphic Aid for Vocational Guidance and Research". In: *Vocational Guidance Quarterly* (1981).

Motivation

- RIASEC labels are usually available only at the occupation level using manual annotation by domain experts
- Such an approach:
  - assumes jobs of the same occupation share the same personality types
  - does not timely profile new occupations

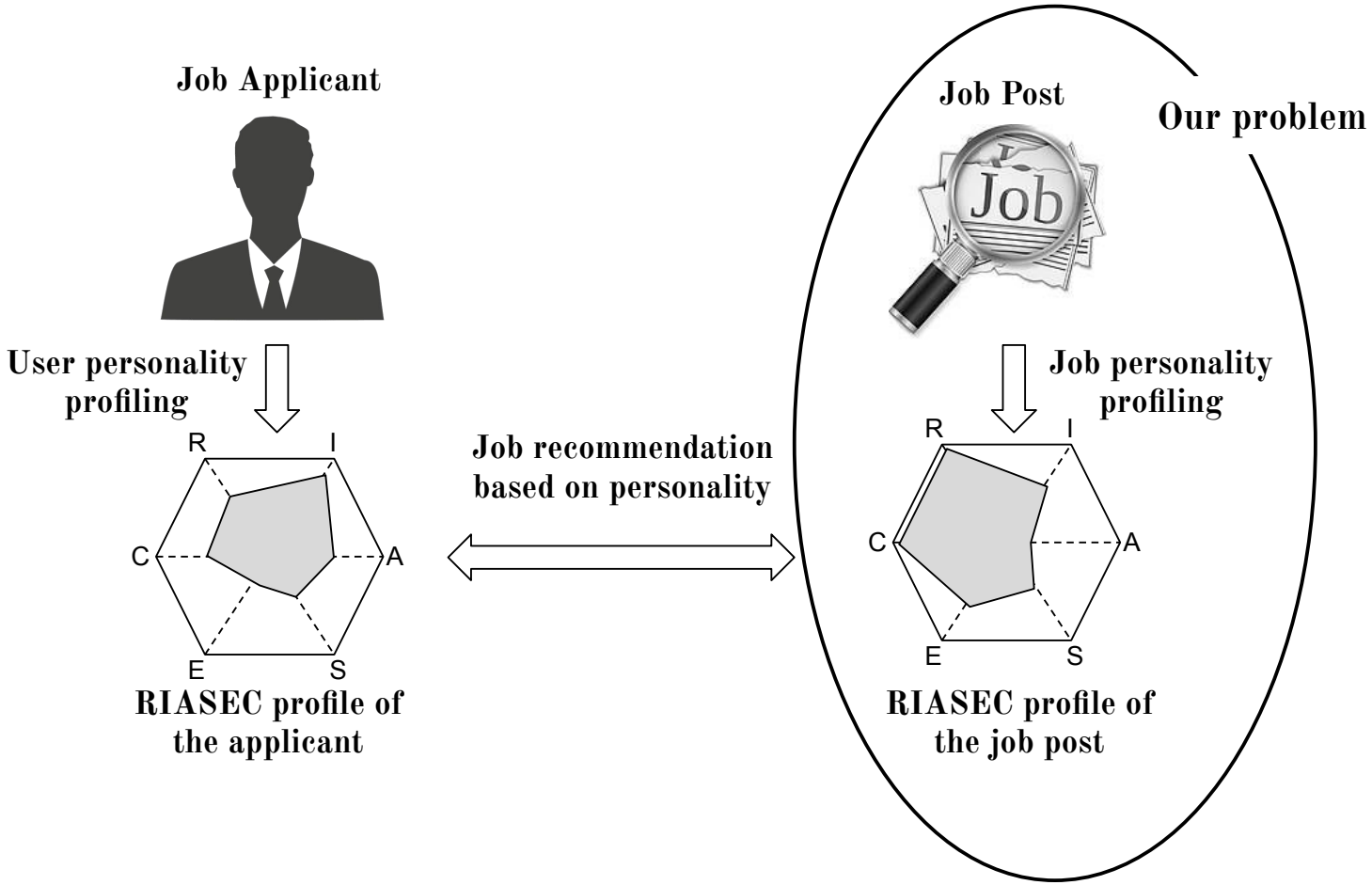
Occupations	Jobs
Web Programmer	Web Architect Web Designer Web Developer
Librarian	Library Director Children's Librarian Library Media Specialist

Figure: A few examples for jobs and occupations (source: O*NET[3])

[3] **https://www.onetonline.org/**.

Problem Statement

- Our aim is to determine the personality types of a large collection of job posts using the text content in job titles and job descriptions



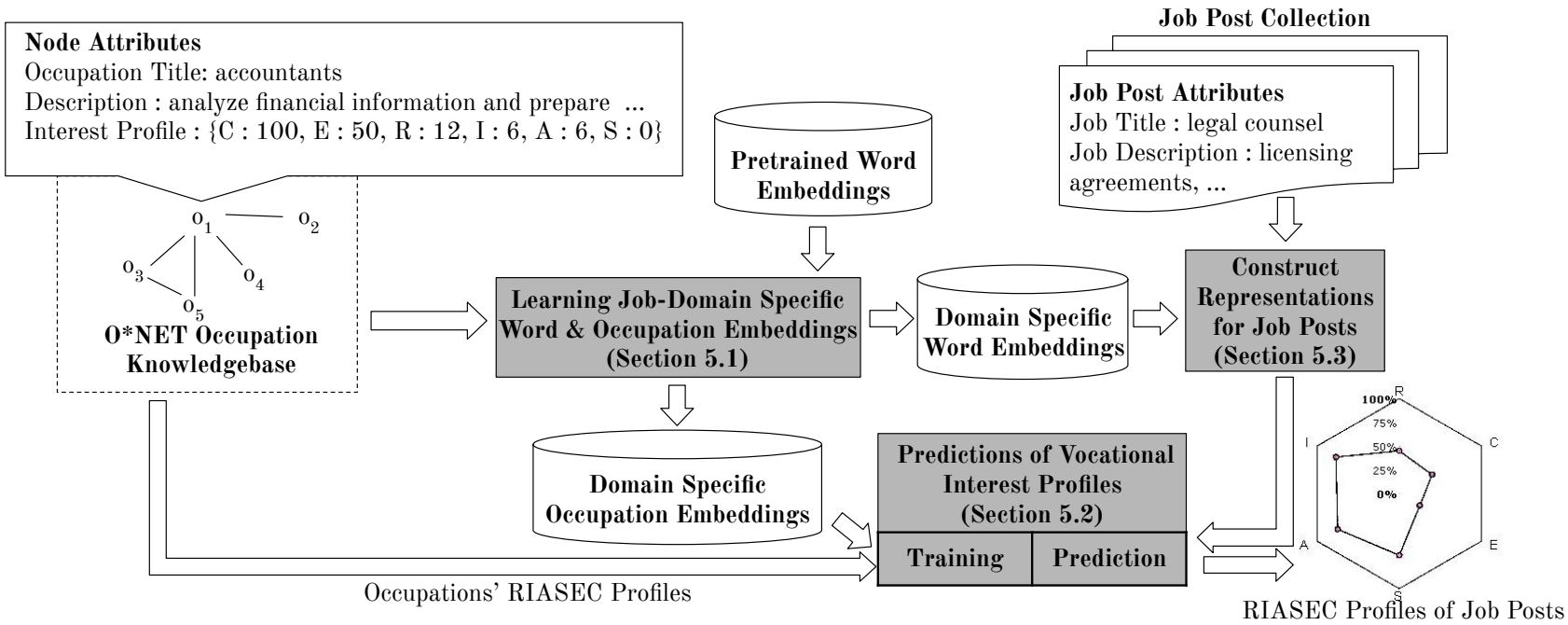
Research Challenges

1. Limited labeled job posts with RIASEC profiles
   How to learn and evaluate a job personality profiling model?
2. Noisy word semantics (e.g., the word "spark" in a software developer job mostly refers to a cluster computing framework, although it means "an emission of fire or electricity" in a general corpus)
   How to capture the job domain-specific word semantics?
3. Ranked RIASEC dimensions for job posts
   How to recover the ranking of RIASEC dimension for a given job post?

Contributions

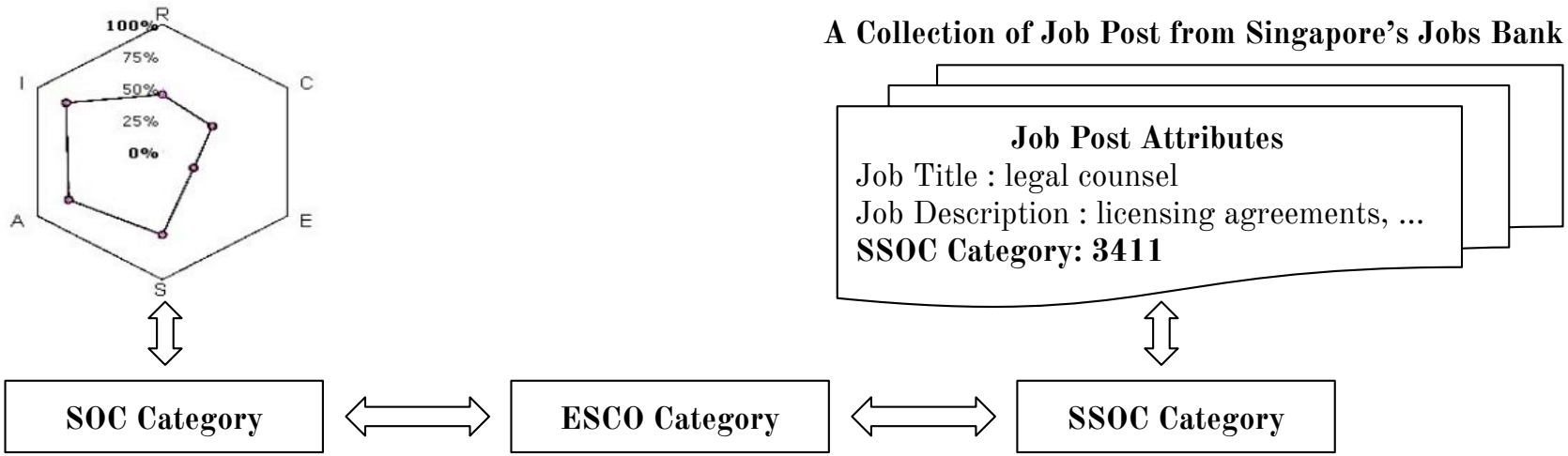
This work proposes:

1. a distant supervision approach to weakly label a collection of job posts
2. an approach to learn job-domain specific word and occupation embeddings to alleviate noisy word semantics
3. a supervised machine learning approach to yield ranked RIASEC dimensions for a given job post based on its text context



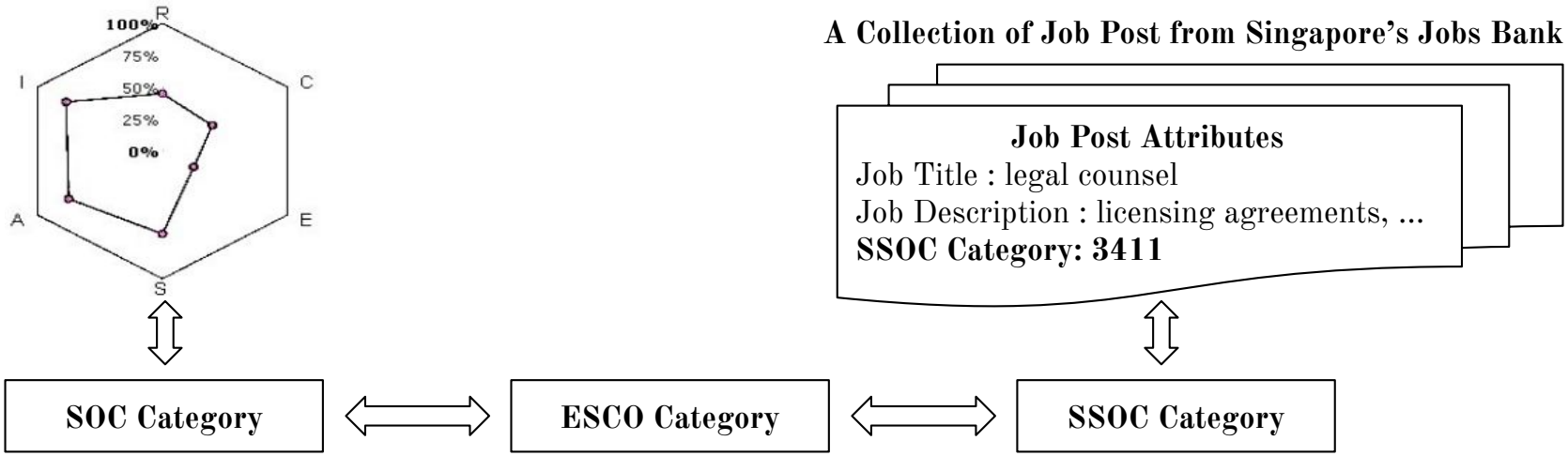
Distant Supervision to Weakly Label Job Posts

- Our source of job posts is Singapore's Jobs Bank, which assigns a SSOC (Singapore Standard Occupational Classification) occupation code to each job post
- The RIASEC profiles are only available for SOC (Standard Occupational Classification) occupation codes, and there is no direct mapping between SOC and SSOC
- In this work, ESCO (i.e., European classification of occupations) occupation codes are used as a bridge to obtain a mapping between SSOC and SOC



Distant Supervision to Weakly Label Job Posts

- The proposed indirect mapping between SSOC and SOC:
  - is able to map 96.71% of SSOC occupations to 96.57% of ESCO occupations, and finally to 75.23% of SOC occupations
  - is used to assign weak labels to the job posts based on the mapped SOC occupation for each job post
  - yields 171,946 job posts (out of 217,874 job posts) with weak RIASEC labels



JPLink-Emb: Learning Job-Domain Specific Occupation and Word Representations

- Occupation-domain Knowledge Graph:
  - occupations and words as nodes
  - occupation-occupation edges are constructed by connecting the similar occupations listed in O*NET
  - occupation-word edges are constructed by connecting keywords in the description of an occupation (in O*NET) for the corresponding occupation
- Breadth-First Search is used on the constructed knowledge graph to generated random-walks
- Word2Vec[4] to jointly learn embeddings for occupations and words using random walks

[4] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Proc. of NIPS*. 2013.

Qualitative Analysis of Representations

Table: 5 most similar occupations and words for the occupation "financial manager", as induced by different embeddings (see our paper for more examples)

	5 most similar occupations			5 most similar words		
Target Occupation	**PretrainedEmb**[5]	**JPLink-Emb**	**Wikipedia2Vec**[6]	**PretrainedEmb**	**JPLink-Emb**	**Wikipedia2Vec**
financial managers	investment fund managers	financial specialists, all other	auditors	managers	investing	brokerage
	security managers	auditors	sales agents, financial services	executives	banking	issuers
	marketing managers	financial managers, branch or department	transportation, storage, and distribution managers	bankers	branch	banking
	logistics managers	treasurers and controllers	financial managers, branch or department	investors	securities	intermediary
	administrative services managers	loan officers	treasurers and controllers	investment	accounting	entities

[5] **https://github.com/mmihaltz/word2vec-GoogleNews-vectors**.

[6] Ikuya Yamada et al. "Joint learning of the embedding of words and entities for named entity disambiguation". In: *Proc. of SIGNLL* (2016).

Construction of Representation for Job Posts

- For each job posts j, the representations are constructed for the job (or occupation) title x_j^t and job (or occupation) description x_j^d by averaging the embeddings of the keywords appear in the title and the description respectively
- The representation of j is computed as:

$$x_j = \beta * x_j^t + (1 - \beta) * x_j^d \tag{1}$$

  where β controls the importance given to title and description in the final representation
- The optimal value for β has been found as 0.6 using a grid search

JPLink-Pred: Learning to Rank Personality Prediction Model

- For a given representation x_o of an occupation (a job), our model predict the scores for RIASEC dimensions $\hat{y}_o$ as:

$$\hat{y}_o = softmax(A * x_o^T + b) \tag{2}$$

where $A_{6\times k}$ and $b_{1\times 6}$ are trainable parameters

- Using the ground truth scores for each occupation (or job) y_o, the modeled top-one probability for each dimension d is computed as:

$$P_d(y_o) = \frac{\exp y_o^d}{\sum_{d' \in \{R,I,A,S,E,C\}} \exp y_o^{d'}} \tag{3}$$

- The following loss function is optimized (motivated by ListNet[7]) using SGD to learn parameters:

$$L(y_o, \hat{y}_o) = \sum_{d \in \{R,I,A,S,E,C\}} P_d(y_o) * \log(\hat{y}_o^d) + (1 - P_d(y_o)) * \log(1 - \hat{y}_o^d) \tag{4}$$

[7] Zhe Cao et al. "Learning to rank: from pairwise approach to listwise approach". In: *Proc. of ICML*. 2007.

Quantitative Results for RIASEC Profile Prediction

Table: RAISEC Profile Prediction Result - NDCG@6 (higher is better)

	Prediction Methods		
Embedding Methods	POINT[8]	PAIR[9]	JPLINK-PRED
PRETRAINEDEMB	0.905	0.912	0.928
WIKIPEDIA2VEC	0.921	0.924	0.941
JPLINK-EMB	**0.928**	**0.934**	**0.949**

- The performance improvements are primarily due to:
  - Learning job-domain specific representation for words and occupations to capture domain specific knowledge
  - Incorporating relative ranking of RIASEC dimensions to capture significant correlations between RIASEC dimensions

[8] **Logistic Regression classifier with binary cross entropy loss**.

[9] Christopher Burges et al. "Learning to rank using gradient descent". In: *Proc. of ICML*. 2005.

Quantitative Results for RIASEC Profile Prediction

Table: Correlations among the dimensions of the linear transformation matrix (A) and the bias values (b) in JPLINK and actual descriptive statistics of O*NET

	Statistics of model parameters							Descriptive statistics of O*NET						
	Correlations between rows of A						b	Actual Correlations						Actual Propotions
	C	E	I	S	R	A		C	E	I	S	R	A	
Conventional (C)	1.00	0.09	-0.11	**-0.29**	0.08	**-0.52**	0.60	1.00	0.21	**-0.28**	**-0.34**	-0.05	**-0.50**	0.24
Enterprising (E)		1.00	**-0.55**	0.12	**-0.31**	**-0.23**	0.12		1.00	**-0.46**	0.15	**-0.51**	-0.06	0.14
Investigative (I)			1.00	**-0.40**	0.10	-0.07	0.15			1.00	-0.09	-0.17	0.11	0.16
Social (S)				1.00	**-0.52**	-0.01	-0.63				1.00	**-0.60**	**0.24**	0.11
Realistic (R)					1.00	**-0.34**	0.65					1.00	**-0.39**	0.28
Artistic (A)						1.00	-0.89						1.00	0.06

- The correlations among the rows of A reflect the actual correlations between the corresponding RIASEC dimensions.
- The bias values (b) of our model are consistent with the actual distribution of RIASEC dimensions in O*NET.

Error Analysis

Table: A few job posts predicted with "wrong" RIASEC profile by JPLink

Assigned SSOC Occupation	Job Title	key words in Job Description	Actual (y) and Predicted ($\hat{y}$) RIASEC Ranking (first has the highest rank)
Graphic and Multimedia Designers and Artists	lead full stack developer	widgets saas javascript html css community	$y : \{A, R, E, I, C, S\}$ $\hat{y} : \{C, I, R, E, S, A\}$
Graphic and Multimedia Designers and Artists	singapore researcher	subject matter experts project managers executive leadership travel presentations oral independent	$y : \{A, R, E, I, C, S\}$ $\hat{y} : \{I, C, R, E, S, A\}$
Other Craft and Related Workers	line leader	coaching	$y : \{R, E, C, A, I, S\}$ $\hat{y} : \{E, C, S, A, R, I\}$
Manufacturing Labourers and Related Workers	on executive	online market place inventory management customer service stocks filing enquires commerce	$y : \{R, C, E, I, A, S\}$ $\hat{y} : \{C, E, I, S, R, A\}$
Information and Communications Technology Installers and Servicers	part time coach	workshops spark pointers learning funds curiosity	$y : \{R, C, I, E, S, A\}$ $\hat{y} : \{E, S, C, I, A, R\}$

- Job titles and descriptions do not tally with the occupations assigned by the current system
- However, the predicted interest profiles are reasonable with respect to their descriptions and titles

Summary

- This work proposed JPLink, a framework to automate the profiling of jobs with their interest profiles, which:
  1. explored the domain-specific knowledge available in O*NET to improve existing word and occupation representations
  2. proposed a novel loss function for the prediction of RIASEC profiles, which captures the interrelationship between RIASEC dimensions
  3. managed to correct a type of imperfections in the current system by profiling job posts at their granular level
- JPlink could be applicable to improve existing job recommendation engines to provide accurate and personalized jobs for applicants

Thank You