Liu *et al. Journal of Big Data* (2022) 9:99
https://doi.org/10.1186/s40537-022-00649-5

Journal of Big Data

RESEARCH Open Access

Title2Vec: a contextual job title embedding for occupational named entity recognition and other applications

Junhua Liu[1,3], Yung Chuen Ng[1], Zitong Gui[1], Trisha Singhal[1], Lucienne T. M. Blessing[1], Kristin L. Wood[1,2] and Kwan Hui Lim[1*]

*Correspondence: kwanhui_lim@sutd.edu.sg

[1] Singapore University of Technology and Design, Singapore, Singapore
[2] University of Colorado Denver, Singapore, Singapore
[3] Forth AI, Singapore, Singapore

Abstract

Occupational data mining and analysis is an important task in understanding today's industry and job market. Various machine learning techniques are proposed and gradually deployed to improve companies' operations for upstream tasks, such as employee churn prediction, career trajectory modelling and automated interview. Job titles analysis and embedding, as the fundamental building blocks, are crucial upstream tasks to address these occupational data mining and analysis problems. A relevant occupational job title dataset is required to accomplish these tasks and towards that effort, we present the Industrial and Professional Occupations Dataset (IPOD). The IPOD dataset contains over 475,073 job titles based on 192,295 user profiles from a major professional networking site. To further facilitate these applications of occupational data mining and analysis, we propose *Title2vec*, a contextual job title vector representation using a bidirectional Language Model approach. To demonstrate the effectiveness of *Title2vec*, we also define an occupational Named Entity Recognition (NER) task and proposed two methods based on Conditional Random Fields (CRF) and bidirectional Long Short-Term Memory with CRF (LSTM-CRF). Using a large occupational job title dataset, experimental results show that both CRF and LSTM-CRF outperform human and baselines in both exact-match accuracy and F1 scores. The dataset and pre-trained embeddings have been made publicly available at https://www.github.com/junhua/ipod.

Keywords: Social computing, Word embedding, Named entity recognition, Occupational mining, Social networks

Introduction

There is a growing interest in occupational data mining and analysis tasks in recent years, especially with the rapid digitization of today's economy and jobs. Furthermore, the advancement of AI and robotics are changing every industry and every sector, drastically altering the employment modes of the work force, especially those with high levels of repetition. Occupational data mining and analysis is also an important research area, which has spurred numerous works in related topics such as employee churn prediction

[1–3], professional career trajectory modelling [4, 5] and predicting employee behaviors with various factors [6, 7], among others.

An important requirement for occupational data mining and analysis is the need for an occupation-related dataset to perform these tasks. Such a dataset could be collected from professional networking sites, derived from online resumes or other sources. While there has been research on these topics, a majority of the datasets used in these earlier research are not publicly available, which further impeds future research in this area. To address this limitation, we present the Industrial and Professional Occupation Dataset (IPOD), which is a publicly available dataset that contains 475,073 job titles/positions belonging to 192,295 users on a major professional networking site. Table 1 shows an overview of relevant datasets and our proposed IPOD dataset, as well as their respective sizes and availability. IPOD is the largest publicly available occupation-related dataset as far as we are aware. We believe that IPOD will be relevant and useful to researchers and industry practitioners who are involved in tasks related to occupational data mining and analysis.

For various occupational data mining and analysis problems, such as next job prediction, churn prediction and others, there is a need to represent the numerous jobs/occupation before using this representation as input to their respective algorithms.

Table 1 A survey of related works that uses and/or provide occupational-related datasets

Publication	Data source	Size	Publicly available
IPOD (our proposed dataset) [10]	**Professional network**	**475K**	**Yes**
Mimno et al. [5]	Resumes	54K	No
Lou et al. [11]	Linkedin	67K	No
Paparrizos et al. [12]	Web	5M	No
Zhang et al. [13]	Job site	7K	No
Liu et al. [4]	Social network	30K	No
Li et al. [14]	Linkedin	–	No
Li et al. [15]	High tech company	–	No
Yang et al. [16]	Resumes	823K	No
Zhu et al. [17]	Job portals	2M	No
James et al. [1]	APS	60K	Yes
Yang et al. [2]	Various channels	–	No
Xu et al. [18]	Professional network	20M	No
Qin et al. [19]	High tech company	1M	No
Lim et al. [20]	Linkedin	10K	No
Shen et al. [21]	High tech company	14K	No
Zhang et al. [22]	Resumes	2.1M	No
Nigam et al. [23]	Job portals	4k	No
Meng et al. [24]	Professional network	414k	No
Van Huynh et al. [25]	Job portals	10k	Yes
Gugnani et al. [26]	Job portals	1.1M	No
Alanoca et al. [27]	Resumes	5k	No
Zhang et al. [28]	Linkedin	459k	No

Apart from two datasets of comprising publications and authors [1] and job titles and descriptions [25], there are no publicly available occupational-related dataset from our survey. The first dataset [1] contains publications and authors from the American Physics Society (APS) but only describes the names and affiliations of physics scientists without their job title or appointments, while the second dataset [25] contains the job title and job description from a job portal but only pertaining to IT-related jobs. Our proposed dataset, IPOD is bolded and in the first row

Many traditional works on occupational data mining and analysis use one-hot encoding or Bag-of-Words to represent job titles, which treats each job title as a distinct entity without being able to model any relationship between them. For example, "Data Scientist" and "Data Analyst" are treat as different jobs using this representation but the two jobs have more similarities than differences. To address this problem, we propose *Title2vec*, a contextual job title vector representation using a bidirectional Language Model approach.

In this paper, we add to this literature of works and make the following contributions:

- To facilitate further research on occupational data mining and analysis, we present and make publicly available the Industrial and Professional Occupation Dataset (IPOD). This dataset consists of 475,073 entries of occupational positions, drafted by working professionals for their user profiles on a major professional networking site, with the motivations of displaying their career achievement, attracting recruiters or expanding professional networks. Out of 23 relevant datasets used in recent works, only three datasets were made publicly available and our proposed IPOD is the largest publicly available dataset among all these used in recent works, as shown in Table 1.
- Using this large occupational dataset of job titles, we performed various occupational data analysis, such as investigating the effects of education on job retention, promotion to management in two distinct cities in US (Denver) and Asia (Singapore).
- To better model job titles and positions, we propose *Title2vec*, a contextual job title vector representation with a bidirectional Language Model [8] approach. This upstream embedding task map the raw job titles into a high-dimensional vector space that allows and boosts the performance of several downstream tasks such as occupational Named Entity Recognition. An example of this occupational NER task is shown in Fig. 1.
- To further demonstrate the usefulness of *Title2vec*, we propose two models for solving this challenging occupational NER task. These two algorithms include a probabilistic machine learning model, namely Conditional Random Field (CRF), and a state-of-the-art recurrent neural network model, namely bidirectional Long Short-Term Memory with CRF (LSTM-CRF) [9]. We compared our algorithms against three baselines and results show that our proposed algorithms out-perform them in terms of Exact Match (EM) accuracy and F1 scores for both overall and tag-specific results.

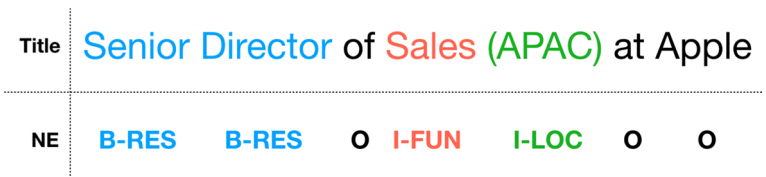

Fig. 1 An example of the occupational NER task, with tokens indicating **RES**ponsibility, **FUN**ction, and **LOC**ation. The NE tags are also added with positional prefixes using BIOES scheme, i.e., **B**egin, **I**nside, **O**thers, **E**nding and **S**ingle

The rest of this paper is structured and organized as follows. Firstly, "Related work" section discusses key works that are related to occupational data mining and analysis, and the use of embedding and named entity recognition techniques for related tasks. Thereafter, "Description of the industrial and professional occupations dataset (IPOD)" section provides an overview of IPOD and describes the key characteristics of this dataset. Using this dataset, "Job title embedding and named entity recognition" section then highlights the approach we proposed for generating job title embedding and describes the occupational name entity recognition task. Following which, "Experiments and results" section outlines our overall experimental methodology for evaluating the proposed algorithm and various baselines. "Occupational data analysis" section shows our occupational data analysis where we investigated retention duration in jobs and time to reach management. Finally, "Conclusion and future work" section summarizes the paper and highlights some future direction for this work.

Related work

In this section, we review related works, in the area of occupational-related datasets, general occupational data mining and analysis, contextual embedding and named entity recognition.

Occupational-related datasets

The recent years have seen numerous research on occupational data mining and analysis related tasks, which utilize some form of occupational-related datasets. To provide an overview of these datasets, we conducted a survey of papers that make use of these datasets and present 23 such papers that were published from 2008 onwards. Table 1 shows our literature review of the related publications that utilizes these types of occupational-related dataset. Out of these 23 datasets, only two been made publicly available, which are [1] by James et al. [25] by Van Huynh et al. and [10] that is described in this paper. The former dataset by James et al. [1] contains the affiliations that physics researchers belong to but does not include their job titles or positions, whereas our dataset comprises the job titles of general workers across the broader industry. Similarly, the dataset by Van Huynh et al. [25] contains 10k job titles and descriptions collected from a job portal but are only relating to jobs within the IT sector.

Named Entity Recognition is another related task to our work and existing NER corpora [29–32] typically use tags, such as **LOC**ation, **PER**son, **ORG**anization, **MISC**ellaneous, etc., that are more application for general tasks. In contrast, IPOD provides domain-specific NE tags to denote the properties of occupations, such as ***RESponsibility***, ***FUNction*** and ***LOCation***. Three experts created a gazetteer based on their expertise in occupational titles and labelled the various named entities in IPOD using this gazetteer. The labelling by the three exports have high reliability with no cases where all three annotators disagree, while inter-rater reliability is similarly high with a Percentage Agreement [33] of 0.853 and Cohen's Kappa [34] of 0.778. We further process the labels by including prefix based on the BIOES tagging scheme [35], i.e., **B**egin, **I**nside, **E**nding, **S**ingle, and **O** indicating that a token belongs to no chunk, indicating the positional features of each token in a title.

Occupational data mining and analysis

Prior works on occupational data mining and analysis aim to accomplish a wide range of tasks, such as Career Modeling and Job Recommendation. In the area of Career Modeling, prior works address downstream tasks including career path modeling [4, 5], career movement prediction [1, 2], job title ranking [18], and employability [36]. In terms of Occupation-related Recommendations, past works focus on analysing Person-Job Fit [17, 19, 21] which commonly aims to suggest employment suitability of candidates for companies, and Job Recommendation [13, 37] which on the other hand provides decision analysis for the job seekers. These works commonly leverage real-world data from different sources, including Linkedin [4, 14], resumes [2, 5], job portals [13, 17] and tech companies [21].

The proposed solutions to these problems are based on different approaches. Most works utilize various machine learning approaches, such as linear classification models [1, 2, 4, 15], generative models [5, 18] and Neural Networks [5, 17, 19]. Some take algorithmic approaches, such as statistical inference [1, 5, 21], Graph-theoretic models [12, 36] and recommmender systems with content-based and collaborative filtering [13, 37]. Some works report their time complexity to be polynomial [4, 15].

Word embedding

Traditional word embedding methods aim to construct word-level vector representations that capture the contextual meaning of words. A Neural Network Language Model (NNLM) was initially proposed with the Continuous Bag of Words (CBoW) model and skip-gram model [38]. These efforts subsequently led to a series of NNLM-based embedding works [39], including the popular Word2Vec model [40]. Pennington et al. [41] proposed Global Vectors for Word Representation (GloVe) that uses a much simpler approach of constructing global vectors to represent contextual knowledge of the vocabulary, which also achieved good results. More recently, a series of high quality contextual models were proposed, such as Embeddings from Language Model (ELMo) [8], FastText [42] and Flair [43]. Both word-level contextualization and character-level features are commonly used for these works and frequently utilized for applications ranging from text classification [44–47] to prediction and recommendation systems [48–51].

Document embedding

While Word Embedding construct static continuous vectors at the word-level, recent works have also proposed methods to represent document-level embeddings. Transformer-based approaches have gained popularity among researchers and shown good results in recent years. Typical applications of these approaches uses pre-trained transformer-based models with very large datasets to construct the document-level embeddings, such as Bidirectional Encoder Representations from Transformers (BERT) [52] and variants of Generative Pre-trained Transformer (GPT) [53], among others. These approaches enables contextual embedding in both word level and document level. Lample et al. [54] proposed a novel Stacked Embedding approach that constructs a hierarchical embedding architecture of document-level embedding by stacking word-level embedding with character-level features and concatenating with an RNN, which

performs well in NER tasks. Apart from document-level embeddings, there are various works that uses Siamese networks and its variants to generate sentence-level embeddings [55, 56].

Named entity recognition

Named entity recognition is a challenging task that traditionally involves a high degree of manually crafted features for different domains. While this task is challenging, the good news is that numerous general large-scale corpora, such as CoNLL-2003 [30] and Ontonotes [31], have been made available for training with deep neural architectures. Popular NER models are based on the LSTM architecture [9, 54], where feeding the sentence embeddings into an uni-directional or bi-directional LSTM encoder. Instead of decoding directly, some works also add a Conditional Random Field (CRF) layer at the end while training the classifier, and use Viterbi [57] to decode the probabilistic output into NE labels. In recent years, the popular transformer-based models [52, 53] have also demonstrated its capability for producing good results.

While manually tagging a large dataset requires tremendous amount of efforts, prior works leverage knowledge-based gazetteers developed by various unsupervised or semi-supervised learning approaches [58, 59], or rely on generative models [60, 61]. Tags can be further formatted with tagging schemes such as IOB [62] or BIOES [35], to indicate the position of tags in a chunk.

Description of the industrial and professional occupations dataset (IPOD)

In this section, we describe our data collection process and characteristics of the IPOD dataset.

Data collection

We obtained over 475,073 job titles based on the user profiles from Asia and the United States, as representatives of the world's most competitive economies [63]. These job titles were based on the career history of 192,295 users with 56.7% from the United States and 43.3% from Asia. After collecting the data, we further pre-processed the data with the standard steps of lowercase conversion, replacement of certain punctuation to words (e.g., substituting & with *and*) and removing special symbols. While lemmatization is also a standard process, we observe that the original unshortened word suggests its most accurate named entity, i.e., Investor is labeled as RESponsibility while Investment is labeled as FUNction, and hence decided not to lemmatize or stem these words.

Description of dataset and tags

Job titles serve as a concise indicator for one's level of responsibility, seniority and scope of work, described with a combination of *responsibility*, *function* and *location*. Table 2 shows examples of the occupational NE tags used in this dataset.

Responsibility, as its name suggests, describes the role and duty of a working professional. As shown in Fig. 1, responsibility may include indicators of managerial levels, such as *director, manager* and *lead*, seniority levels, such as *vice, assistant* and *associate*, and operational role, such as *engineer, accountant* and *technician*. A combination of the three sub-categories draws the full picture of one's responsibility.

Table 2 Examples of occupational NE tags. Modified from [10]

Responsibility	Managerial level:
	lead, supervisor, manager, director, president
	Operational role:
	engineer, designer, accountant, technician
	Seniority:
	junior, vice, associate, assistant, senior
Function	Departments:
	sales, marketing, finance, operations, strategy
	Scope:
	enterprise, project, customer, national, site
	Content:
	data, r&d, security, training, integration, education
Location	Regions:
	APAC, SEA, Asia, European, north, central
	Countries/States/Cities:
	China, America, Singapore, Colorado

Function describes business functions in various dimensions. Specifically, *Departments* describes the company's departments the staffers are in, such as *sales, marketing* and *operations*; *Scope* indicates one's scope of work, such as *enterprise, project* and *national*; lastly, *Content* indicates one's content of work, such as *data, r&d* and *security*.

Finally, **Location** indicates the geographic scope that the title owner is responsible of. Examples of this NE tag include geographic regions such as *APAC, Asia, European*, and counties/states/cities such as *China, America* and *Colorado*.

Annotation process

The IPOD dataset defines the occupational domain-specific NE tags as *RES, FUN, LOC* and *O*, indicating the *responsibility, function, location* and *others* respectively. The labels are further processed by adding prefix using BIOES tagging scheme [35], i.e., **B**egin, **I**nside, **E**nding, **S**ingle, and **O** indicating that a token belongs to no chunk, indicating the positional features of each token in a title. For instance, a job title of *chief financial officer Asia Pacific* is tagged as *S-RES S-FUN S-RES B-LOC E-LOC* with the BIOES scheme [35].

We adopt a knowledge-based NE tagging strategy by creating a gazetteer of word tokens, with the help of three experts including a HR personnel, a senior recruiter and a seasoned business professional. These three experts developed the comprehensive gazetteer based on their expertise in occupational titles and subsequently labelled the various named entities in IPOD using this gazetteer. This process includes first running a unigram analysis of the job titles to retrieve the top 1,500 tokens in terms of occurrence frequency of the uni-grams, which are subsequently tagged by the three expert annotators. The distribution of IPOD based on these labels are presented in Table 3.

Among the 1500 tokens tagged, every tag is agreed with at least two annotators, where 1169 (77.9%) are commonly agreed among all three annotators, and 331 (22.1%) are agreed with two annotators. Of particular note, there are no instances where all three

Table 3 Frequency count of named entities in IPOD

	***RES*ponsibility**	***FUN*ction**	***LOC*ation**	***O*thers**
Frequency count	310,570	255,974	9,998	66,948

annotators totally disagree, i.e., three different labels by each annotator. We further evaluate the Inter-Rater Reliability with two inter-coder agreements, observing a score of 0.853 based on Percentage Agreement [33] and 0.778 based on Cohen's Kappa [34], which reflects a high level of agreement. Finally, the job titles are labelled with NE tags using BIOES scheme and formatted for NER tasks.

Job title embedding and named entity recognition

In this section, we describe our proposed job title vector representation called Title2vec and demonstrate its usefulness for an occupational NER task.

Job title embedding

A fundamental aspect of occupational data mining and analysis is to properly represent job titles, where standard one-hot encoding or Bag-of-Words are unable to accurately capture the relationship between different job titles. Towards this effort, we propose a contextual job title vector representation model, *Title2vec*. *Title2vec* is based on a bidirectional Language Model approach [8] where each token is represented with a contextual vector with 3072 dimensions, before passing through a bi-directional LSTM network. The forward LSTM predicts the probability of each token given its history, and the backward LSTM does the same in reverse order. Compared to uni-directional models, this approach enables us to better capture the meaning of a word in the context of its entire sentence or preceding and subsequent words.

Instead of training from scratch, we construct *Title2vec* by fine-tuning from a pre-trained model, namely the Embedding from Language Models (ELMo) [8]. The choice of ELMo is because it provides a language-level contextual meaning for word tokens that is highly similar, if not identical, to that in job titles. For instance, the word *director* appearing in a job title is similar to its counterpart that is mentioned in a Wikipedia article. Such a contextual language model also better captures the meaning of a word than its non-contextual counterpart. For example, the word "bank" has a different meaning in "senior bank officer" and "bank of the river", which contextual embeddings are better able to represent.

Occupational named entity recognition

To demonstrate the effectiveness of *Title2vec*, we define an occupational Named Entity Recognition task. Unlike typical NER tasks that uses more general tags like **LOC**ation, **PER**son, **ORG**anization, **MISC**ellaneous, etc., the tags for the occupational NER tasks are more specific and relevant to job positions and titles. These occupation-specific tags also require a specialised dataset, which our proposed IPOD dataset is able to satisfy. Figure 1 illustrates an example of this occupational NER task, where the job title of "Senior Director of Sales (APAC) at Apple" is tagged with the named entities of B-RES, B-RES, O, I-FUN, I-LOC, O and O. Regarding these named entities, they refer to the

RESponsibility, **FUN**ction, and **LOC**ation associated with the job title, with a further tagging of **B**eginning, **I**nside, **O**thers, **E**nding and **S**ingle based on the BIOES scheme.

Occupational NER models and baselines

We propose two algorithms to address this occupational NER task. The first algorithm is based on Conditional Random Fields (CRF), which is a probabilistic machine learning model that produce joint probability of the co-occurrence of output sequence [64]. Our second algorithm belongs to the recurrent neural networks family, namely a bidirectional Long Short-Term Memory model with a CRF layer adding to the output layer (LSTM-CRF), as shown in Fig. 2. Variations of the LSTM-CRF model showed state-of-the-art results in some recent works for different downstream NLP tasks [9, 54]. Both the CRF and LSTM-CRF are decoded using a first-order Viterbi algorithm [57] that finds the sequence of NE tags with highest scores. As baselines for subsequent evaluation, we construct two baseline encoders based on a Logistic Regression (LogReg) classifier and a standard LSTM, both of which are decoded using a softmax layer. We also use a human annotation as our third baseline to highlight how these algorithms perform compared to a manual process.

Hyper-parameter optimization

To ensure an optimal selection of parameters, we conduct a grid search of hyper-parameters to fine-tune the performance of our proposed CRF and LSTM-CRF models. The search space includes varying learning rates (i.e., 0.1 or 0.01), number of LSTM hidden layers (1 or 2), number of hidden states (128 or 256), mini-batch size (32 or 128) and type of optimizer (Adam or SGD). We use Word Dropout and Variational Dropout [65] to prevent over-fitting, with probability of 0.05 and 0.5 respectively. In total, we perform a comprehensive evaluation of over 100 hyper-parameters sets, where each set of hyper-parameters is run with 10 epochs.

Table 4 shows the breakdown of the search space and the final hyper-parameters used for both models. We deploy a Cross Entropy loss function and a SGD optimizer with an initial learning rate of 0.1 and a mini-batch size of 32 for both proposed CRF model and baselines. For the two LSTM-based models, we use a single hidden layer with an initial learning rate of 0.1, LSTM state size of 256 and a mini-batch size of 128.

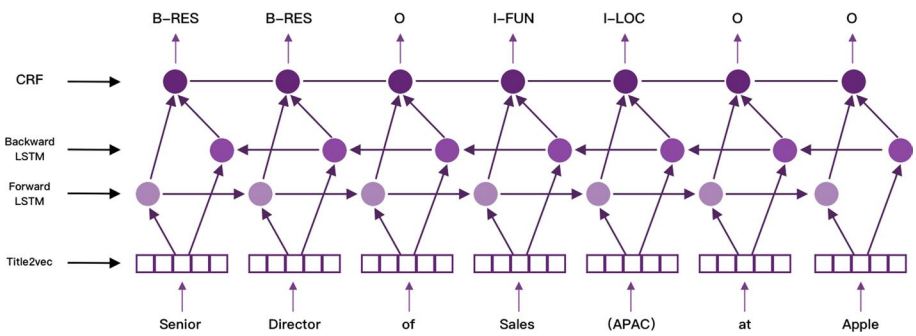

Fig. 2 Bidirectional LSTM-CRF model for occupational NER task

Table 4 Hyper-parameter search space and final values used

Hyper parameter	CRF		LSTM-CRF	
	Final	Range	Final	Range
Learning rate	0.1	{0.01, 0.1}	0.1	{0.01, 0.1}
Mini-batch size	32	{32, 128}	128	{32, 128}
Word dropout	0.05	–	0.05	–
Variational dropout	0.5	–	0.5	–
Type of optimizer	SGD	{Adam, SGD}	SGD	{Adam, SGD}
LSTM layers	–	–	1	{1, 2}
LSTM state size	–	–	256	{128, 256}

Table 5 Comparison of results for occupational NER

Models	Precision	Recall	Exact match	F1-score
LogReg	90.80	93.20	85.10	92.00
LSTM	99.71	99.90	99.61	99.80
Human	91.60	99.60	91.30	95.40
CRF	**99.90**	99.81	99.71	99.85
LSTM-CRF	99.86	**99.97**	**99.83**	**99.91**

Bolded text is used to indicate the best performing values

Experiments and results

In this section, we discuss our experimental evaluation and results.

Evaluation metrics

For evaluating the performance of our proposed algorithms against the three baselines, we use the standard metrics of *Precision*, *Recall*, *Exact Match (EM)* and the *F1 score*. The *F1* score, formally defined as $F1 = 2 * Precision * Recall * (Precision + Recall)^{-1}$, allows us to measure the average overlaps between the ground truth and predicted tags. The *EM* metric measures the percentage agreement between the ground truth and predicted tags with exact matches. These metrics are commonly used in various NLP tasks and allows us to evaluate the performance of our proposed algorithms against the various baselines.

Human performance

We construct the human performance baseline for IPOD using the NE tags annotated by the three domain experts. We choose the set of labels tagged by annotator 1 as the ground truth labels and compute against the other two annotation sets. We then take the average *EM* and *F1* to indicate human performance. We record an *EM* accuracy of 91.3%, and an *F1* of 95.4%. This shows a strong human performance as compared to those of other datasets, such as 91% *EM* for the CHEMDNER corpus [66], 86.8% EM and 89.5% F1 for SQuAD2.0 [67], and 77.0% *EM* and 86.8 *F1* for SQuAD1.0 [68].

Table 6 Comparison of results stratified by NE Tags

	FUN		LOC		RES	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
LogReg	78.30	87.80	93.70	96.80	90.10	94.80
LSTM	99.49	99.74	97.68	98.83	99.77	99.88
CRF	99.35	99.67	**98.96**	**99.48**	**99.99**	**99.99**
LSTM-CRF	**99.88**	**99.94**	98.70	99.35	99.82	99.91

Bolded text is used to indicate the best performing values

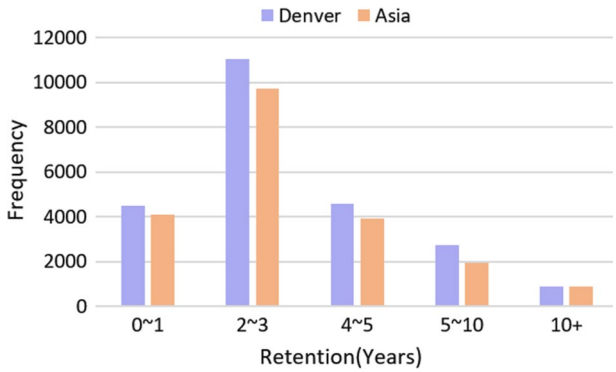

Fig. 3 Retention rate of users in Denver and Asia in our dataset

Experimental results

Table 5 shows the overall performance of our proposed CRF and LSTM-CRF models against the three baselines, in terms of precision (P), recall (R), exact match (EM) and F1 score. In all cases, our proposed CRF and LSTM-CRF models out-perform all baselines in terms of all metrics. When comparing between CRF and LSTM-CRF, we observe that the latter performs better overall with a slight under-performance in terms of precision. Interestingly, the results also show that the Human baseline out-performs the LogReg baseline.

For a more detailed analysis, we now examine the performance of the various algorithms on a per-tag basis. Table 6 shows the per-tag breakdown of the NER results, in terms of EM and F1. Our proposed CRF and LSTM-CRF algorithms offer similar levels of performance, and consistently out-perform both the LogReg and LSTM baselines across all three categories of tags. These results highlight the effectiveness of our proposed model against the baselines as well as the advantages over manual annotation (via the human baseline).

Occupational data analysis

In addition to proposing the *Title2vec* model for job title embedding and evaluating it on the occupational NER task, we also perform a preliminary exploratory data analysis on the same dataset to uncover trends related to occupational data mining and analysis. In particular, we examine the retention duration in jobs and the time to reach management positions.

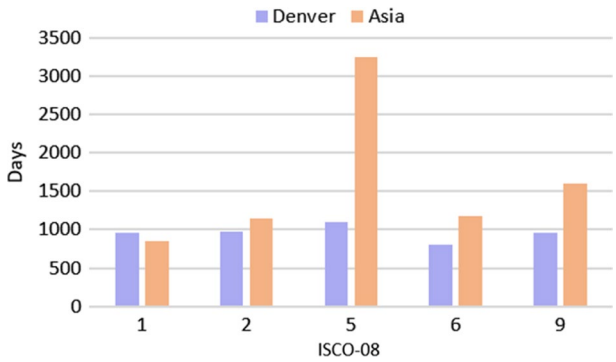

Fig. 4 Time to reach management for users in our dataset

Table 7 Major groups of jobs based on ISCO-08 classification

ISOC-08 code	Occupation group
1	Managers
2	Professionals
3	Technicians and associate professionals
4	Clerical support workers
5	Services and sales workers
6	Skilled agricultural, forestry and fishery workers
7	Craft and related trades workers
8	Plant and machine operators and assemblers
9	Elementary occupations
0	Armed forces occupations

Retention duration in jobs

We first examine the number of years a user stays in a single job position, i.e., the retention rate of a job. Figure 3 shows the retention rate (in years) of users in US (Denver) and Asia (Singapore). In both regions, we can see that most users stay between 2 and 3 years in a single job, while the least number of users stay beyond 10 years in the same job. In general, there is a higher number of longer retention in jobs in Denver than in Asia, with no noticeable difference for those beyond 10 years.

Next, we investigate the hypothesis that the duration of education of a user can be used to predict the retention of the user in a job. We perform a simple linear regression analysis to determine if the number of years of education is able to predict how long a user will stay in a job. The analysis returned results of $R^2 = 0.0008, p = 0.0016$ for Asia and $R^2 < 0.0001, p = 0.2668$ for Denver. This shows the amount of education in terms of duration significantly predicted employee retention for Asia ($b = -0.9422, p = 0.0016$), but there is no sufficient evidence to indicate that for Denver ($b = -1.795, p = 0.2668$).

Time to reach management position

Figure 4 illustrates the number of years users take to reach a management level position in Denver and Asia. The users are further split by the International Standard Classification of Occupations ISCO-08 codes of their job titles, which is a classification of

different groupings of jobs based on their task and functions. Table 7 shows examples of the major groups of jobs based on ISCO-08.

Across most job categories, users in Asia take a longer time to reach a management level position than their counterparts in Denver. This trend is especially pronounced for those in job category 5 (Services and Sales Workers), with users in Asia taking three times longer than those in Denver. This observation could potentially be due to differing industry distributions and working culture in the two different regions.

Similar to the previous section, we now study the hypothesis of whether the number of years of education that a user has undertaken can be used to predict the time needed for him/her to reach a management position. To investigate this hypothesis, we conduct a simple linear regression analysis to determine if a user's education duration (in years) is able to predict how long a user will take to reach a management position. This analysis produced results of $R^2 = 0.0002, p = 0.5570$ for Asia and $R^2 = 0.0010, p = 0.0004$ for Denver. The results highlight that the duration of education is able to significantly predict the time needed to reach a management position for users in Denver ($b = 9.325, p = .0004$), but is insufficient to conclude the same for users in Asia ($b = 6.382, p = 0.5570$).

Conclusion and future work

In this paper, we explored the general research area of occupational data mining and analysis in terms of providing a relevant occupational-related dataset and the related tasks of job title embeddings and occupational NER. We presented the IPOD corpus that consists of 475,073 job titles from 192,295 users on a major professional networking site, with a knowledge-based gazetteer that includes manual NE tags from three domain exports annotators. As far as we are aware, IPOD is the largest publicly available dataset about occupational titles across the general industry.

We also addressed two challenging upstream tasks of occupational data mining and analysis, namely job title embeddings and occupational NER. Towards addressing these tasks, we proposed *Title2vec*, a contextual job title vector representation using a bidirectional Language Model approach and developed the CRF and bidirectional LSTM-CRF models for solving the important occupational Named Entity Recognition problem. Experimental results on the IPOD dataset show that our proposed CRF and bidirectional LSTM-CRF models out-perform the various baselines in terms of exact-match accuracy and F1 scores, for the overall NER task as well as on a per-tag basis. We also performed a preliminary exploratory data analysis of this dataset, examining the retention duration and time to reach management positions for these users.

For future work, there are various possible directions to explore. Firstly, we could explore BERT-based models for performing similar occupational NER tasks. However, we note that such BERT-based models would be more complex in terms of number of parameters than the current models examined in this paper. Moreover, our proposed LSTM-CRF model already achieved scores of more than 99.8% in terms of Precision, Recall, F1-score and Exact Match, thus any further improvements would not be relatively minor. Secondly, the job title embeddings generated by this model could be used to facilitate various downstream tasks, such as predicting job churning and recommending

future job positions. Lastly, we can further expand on our preliminary exploratory data analysis by exploring how other user attributes (e.g., education level, discipline of study, work experience, etc.) affect retention duration and time to reach management.

Acknowledgements
This research is funded in part by the Singapore University of Technology and Design under grant SRG-ISTD-2018-140.

Author contributions
JL designed the research, ran experiments, analyzed results, and wrote the manuscript. YCN and ZG designed the research, ran experiments, analyzed results and contributed to manuscript preparation. TS, LM, KW and KHL designed the research, analyzed results and contributed to manuscript preparation. All authors read and approved the final manuscript.

Funding
This research is funded by SRG-ISTD-2018-140.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors consent to the publication of this manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 22 November 2021 Accepted: 7 August 2022
Published online: 03 September 2022

References

1. James C, Pappalardo L, Sîrbu A, Simini F. Prediction of next career moves from scientific profiles. arXiv preprint. 2018. arXiv:1802.04830.
2. Yang Y, Zhan D-C, Jiang Y. Which one will be next? An analysis of talent demission. 2018.
3. Zhao Y, Hryniewicki MK, Cheng F, Fu B, Zhu X. Employee turnover prediction with machine learning: a reliable approach. In: Proceedings of SAI intelligent systems conference. Springer;2018. p. 737–58.
4. Liu Y, Zhang L, Nie L, Yan Y, Rosenblum DS. Fortune teller: predicting your career path. In: Thirtieth AAAI conference on artificial intelligence. 2016.
5. Mimno D, McCallum A. Modeling career path trajectories. Citeseer; 2008.
6. Chen Z. Mining individual behavior pattern based on significant locations and spatial trajectories. In: 2012 IEEE international conference on pervasive computing and communications workshops. IEEE;2012. p. 540–1.
7. Cetintas S, Rogati M, Si L, Fang Y. Identifying similar people in professional social networks with discriminative probabilistic models. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. 2011. p. 1209–10.
8. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. arXiv preprint. 2018. arXiv:1802.05365.
9. Liu L, Shang J, Ren X, Xu FF, Gui H, Peng J, Han J. Empower sequence labeling with task-aware neural language model. In: Thirty-second AAAI conference on artificial intelligence; 2018.
10. Liu J, Ng YC, Wood KL, Lim KH. IPOD: a large-scale industrial and professional occupation dataset. In: Conference companion publication of the 2020 on computer supported cooperative work and social computing. 2020. p. 323–8.
11. Lou Y, Ren R, Zhao Y. A machine learning approach for future career planning. Citeseer, Technical report; 2010.
12. Paparrizos I, Cambazoglu BB, Gionis A. Machine learned job recommendation. In: Proceedings of the fifth ACM conference on recommender systems. ACM; 2011. p. 325–8.
13. Zhang Y, Yang C, Niu Z. A research of job recommendation system based on collaborative filtering. In: 2014 seventh international symposium on computational intelligence and design, vol. 1. IEEE; 2014. p. 533–8.
14. Li L, Jing H, Tong H, Yang J, He Q, Chen B-C. Nemo: next career move prediction with contextual embedding. In: Proceedings of the 26th international conference on world wide web companion. International World Wide Web Conferences Steering Committee; 2017. p. 505–13.
15. Li H, Ge Y, Zhu H, Xiong H, Zhao H. Prospecting the career development of talents: a survival analysis perspective. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2017. p. 917–25.
16. Yang S, Korayem M, AlJadda K, Grainger T, Natarajan S. Combining content-based and collaborative filtering for job recommendation system: a cost-sensitive statistical relational learning approach. Knowl Based Syst. 2017;136:37–45.

17. Zhu C, Zhu H, Xiong H, Ma C, Xie F, Ding P, Li P. Person-job fit: adapting the right talent for the right job with joint representation learning. ACM Trans Manag Inf Syst (TMIS). 2018;9(3):12.
18. Xu H, Yu Z, Guo B, Teng M, Xiong H. Extracting job title hierarchy from career trajectories: a bayesian perspective. In: IJCAI. 2018. p. 3599–605.
19. Qin C, Zhu H, Xu T, Zhu C, Jiang L, Chen E, Xiong H. Enhancing person-job fit for talent recruitment: an ability-aware neural network approach. In: The 41st international ACM SIGIR conference on research & development in information retrieval. ACM; 2018. p. 25–34.
20. Lim E-P, Lee W-C, Tian Y, Hung C-C. Are you on the right track? Learning career tracks for job movement analysis. In: Workshop on data science for human capital management (DSHCM2018). DSHCM; 2018. p. 1–16.
21. Shen D, Zhu H, Zhu C, Xu T, Ma C, Xiong H. A joint learning approach to intelligent job interview assessment. In: IJCAI. 2018. p. 3542–8.
22. Zhang L, Zhu H, Xu T, Zhu C, Qin C, Xiong H, Chen E. Large-scale talent flow forecast with dynamic latent factor model. In: The world wide web conference. 2019. p. 2312–22.
23. Nigam A, Roy A, Singh H, Waila H. Job recommendation through progression of job selection. In: 2019 IEEE 6th international conference on cloud computing and intelligence systems (CCIS). IEEE; 2019. p. 212–6.
24. Meng Q, Zhu H, Xiao K, Zhang L, Xiong H. A hierarchical career-path-aware neural network for job mobility prediction. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019. p. 14–24.
25. Van Huynh T, Van Nguyen K, Nguyen NL-T, Nguyen AG-T. Job prediction: from deep neural network models to applications. In: 2020 RIVF international conference on computing and communication technologies (RIVF). IEEE; 2020. p. 1–6.
26. Gugnani A, Misra H. Implicit skills extraction using document embedding and its use in job recommendation. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34. 2020. p. 13286–93.
27. Alanoca HA, Vidal AA, Saire JEC. Curriculum vitae recommendation based on text mining. arXiv preprint. 2020. arXiv:2007.11053.
28. Zhang L, Zhou D, Zhu H, Xu T, Zha R, Chen E, Xiong H. Attentive heterogeneous graph embedding for job mobility prediction. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021. p. 2192–201.
29. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics; 2005. p. 363–70.
30. Sang EF, De Meulder F. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. 2003. arXiv:cs/0306050.
31. Weischedel R, Palmer M, Marcus M, Hovy E, Pradhan S, Ramshaw L, Xue N, Taylor A, Kaufman J, Franchini M, et al. Ontonotes release 5.0 ldc2013t19. Philadelphia: Linguistic Data Consortium. 2013. p. 23.
32. Borchmann Ł, Gretkowski A, Gralinski F. Approaching nested named entity recognition with parallel LSTM-CRFs. In: Proceedings of the PolEval 2018 workshop. 2018. p. 63.
33. Viera AJ, Garrett JM, et al. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
34. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. Comput Linguist. 2008;34(4):555–96.
35. Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the thirteenth conference on computational natural language learning. CoNLL '09. Stroudsburg: Association for Computational Linguistics. 2009. p. 147–55. http://dl.acm.org/citation.cfm?id=1596374.1596399.
36. Massoni S, Olteanu M, Rousset P. Career-path analysis using optimal matching and self-organizing maps. In: International workshop on self-organizing maps. Springer; 2009. p. 154–62.
37. Malinowski J, Keim T, Wendt O, Weitzel T. Matching people and jobs: a bilateral recommendation approach. In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06), vol. 6. IEEE; 2006. p. 137.
38. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. J Mach Learn Res. 2003;3(Feb):1137–55.
39. Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In: Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics; 2010. p. 384–94.
40. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. 2013. arXiv:1301.3781.
41. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. p. 1532–43.
42. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–46.
43. Akbik A, Blythe D, Vollgraf R. Contextual string embeddings for sequence labeling. In: Proceedings of the 27th international conference on computational linguistics. 2018. p. 1638–49.
44. Liu J, Singhal T, Blessing LTM, Wood KL, Lim KH. Crisisbert: a robust transformer for crisis classification and contextual crisis embedding. In: Proceedings of the 32nd ACM conference on hypertext and social media (HT'21). 2021. p. 133–41.
45. Singhal T, Liu J, Blessing LT, Lim KH. Analyzing scientific publications using domain-specific word embedding and topic modelling. In: 2021 IEEE international conference on big data (Big Data). IEEE; 2021. p. 4965–73.
46. Kumar S, Zymbler M. A machine learning approach to analyze customer satisfaction from airline tweets. J Big Data. 2019;6(1):1–16.
47. Li M, Lim KH. Geotagging social media posts to landmarks using hierarchical BERT (student abstract). In: Proceedings of the thirty-sixth AAAI conference on artificial intelligence (AAAI'22). 2022.
48. Solanki P, Harwood A, et al. User identification across social networking sites using user profiles and posting patterns. In: 2021 international joint conference on neural networks (IJCNN). IEEE; 2021. p. 1–8.

49. Pek YN, Lim KH. Identifying and understanding business trends using topic models with word embedding. In: Proceedings of the 2019 IEEE international conference on big data (BigData'19). 2019. p. 6177–9.
50. Ho NL, Lim KH. User preferential tour recommendation based on POI-embedding methods. In: Proceedings of the 26th international conference on intelligent user interfaces companion (IUI'21). 2021. p. 46–8.
51. Mu W, Lim KH, Liu J, Karunasekera S, Falzon L, Harwood A. A clustering-based topic model using word networks and word embeddings. J Big Data. 2022;9(1):1–38.
52. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint. 2018. arXiv:1810.04805.
53. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
54. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. arXiv preprint. 2016. arXiv:1603.01360.
55. Reimers N, Gurevych I, Reimers N, Gurevych I, Thakur N, Reimers N, Daxenberger J, Gurevych I, Reimers N, Gurevych I, et al. Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2019.
56. Zhang Y, He R, Liu Z, Lim KH, Bing L. An unsupervised sentence embedding method by mutual information maximization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20). 2020. p. 1601–10.
57. Forney GD. The Viterbi algorithm. Proc IEEE. 1973;61(3):268–78.
58. Kazama J, Torisawa K. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In: Proceedings of ACL-08: HLT. 2008. p. 407–15.
59. Saha SK, Sarkar S, Mitra P. Gazetteer preparation for named entity recognition in Indian languages. In: Proceedings of the 6th workshop on asian language resources. 2008.
60. Nallapati R, Surdeanu M, Manning C. Blind domain transfer for named entity recognition using generative latent topic models. In: Proceedings of the NIPS 2010 workshop on transfer learning via rich generative models. 2010. p. 281–9.
61. Mukund S, Srihari RK. Ne tagging for Urdu based on bootstrap POS learning. In: Proceedings of the third international workshop on cross lingual information access: addressing the information need of multilingual societies. Association for Computational Linguistics; 2009. p. 61–9.
62. Ramshaw LA, Marcus MP. Text chunking using transformation-based learning. In: Natural language processing using very large corpora. Springer; 1999. p. 157–76.
63. Akhtar A. Singapore and Hong Kong have overtaken the US as the most competitive economies. Here's how 25 countries rank. Business Insider. 2019. https://www.businessinsider.com/most-competitive-economies-in-the-world-2019-5.
64. Lafferty J, McCallum A, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. 2001.
65. Kingma DP, Salimans T, Welling M. Variational dropout and the local reparameterization trick. In: Advances in neural information processing systems. 2015. p. 2575–83.
66. Martin K, Obdulia R, Florian L, Miguel V, David S, Zhiyong L, Robert L, Yanan L, Donghong J, Lowe DM. The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform. 2015;7(S1):2.
67. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for squad. In: Proceedings of the 56th annual meeting of the association for computational linguistics, vol. 2 (short papers). 2018. https://doi.org/10.18653/v1/p18-2124.
68. Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 conference on empirical methods in natural language processing. 2016. https://doi.org/10.18653/v1/d16-1264.

Publisher's Note