Improve Job Experience Prediction with Attention Mechanism

Truong Dinh Thien*, Ton Quang Hoang Nguyen

Abstract—Employee in information technology domain is active in finding new job to advance their career path. In that case, company will deal with human resource exhaustion if employee quit their current job. Understanding their job mobility can benefit the company in a variety of ways. While most studies focus on predicting next job title, the problem of forecasting working duration of employees at individual level receives little attention. Moreover, previous methods treat different experiences as similarly important so they cannot utilize potential connections between experiences. To solve the above problems, we contribute new model with attention mechanism. In particular, attention mechanisms give more understanding to learned representations. Different from previous works, our model can effectively utilize previous employee experiences and flexibly adapts to the information of different importance. Our method is applied for 10.000 real-world profiles and shows significant results that outperform the strong baseline model and other state-of-the-art methods.

Index Terms—Attention Mechanism; Career Path; Job Duration.

1. Introduction

It is normal that employees in the company will search for more motivation in their careers. Especially in the IT domain, the employee can change their job title rapidly and have a significantly high turnover rate compared to other domains. To prevent this phenomenon, understanding the pattern of employee mobility is a convenient approach and can bring multiple benefits to the company. Predicting the next job title and estimating the time an employee works in a new job title are important for employee management in the company. Executives can make promote decisions at the right time to reduce the turnover rate and effectively leverage employee talent.

So far, traditional methods such as surveys [1], data mining [2] have been applied to understand job mobility at the organization level. However, these methods have some limitations in investigating employee moving patterns at the individual level and they also require a lot of sensitive data that is very hard to collect from online websites.

In the case of the next job title prediction, job recommender systems [3] based on traditional approaches such as Content-Based Filtering and Collaborative Based Filtering [4] did not give desired results. The sparsity of Job-Person interaction matrices limits the performance of collaborative-based methods. When the information contained in job contents is insufficient, or the feature-engineering on contents is difficult, content-based approaches may become inaccurate.

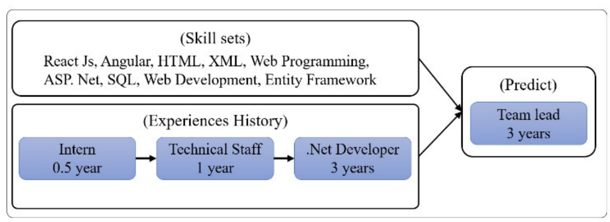

Fig. 1: Sample data and prediction

Furthermore, these methods generally lack the ability to model personal information. To address the problem of job transition, some works are proposed, Markov Chain sequential model [5] that embeds each company and title into the semantic space and considers only the previous company/title in the prediction phase. Deep learning methods were applied to solve this problem. Van Huynh T. et al use the IT job description collected from the online finding-job sites to predict job title [6]. Padmaja Pulicherla et al try to predict job shifting of employee by using employee's information such as gender, salary, company size, and working hours [7]. These approaches can only apply to the organization level and it's hard to collect sensitive information while not giving insight information to benefit the back office. Other method achieves higher applicability by using the experience to predict next job title. NEMO [8] explores action dependencies in career paths and transforms them into contextual embedding to predict the next career move. A hierarchical Long Short-Term Memory (LSTM) is used to predict the next potential employer of

Truong Dinh Thien is with Computer Science and Engineering Department, Neurond, Orient Da Nang, Viet Nam (E-mail: thien.truong@orientsoftware.com).

Ton Quang Hoang Nguyen is with Computer Science and Engineering Department, Neurond, Orient Da Nang, Viet Nam.

**Corresponding author: Truong Dinh Thien (E-mail: thien.truong@orientsoftware.com)*
Manuscript received March 04, 2022; revised May 10, 2022; accepted June 06, 2022.
Digital Object Identifier 10.31130/ud-jst.2022.105E

a person and how long he/she will stay in the new position [9]. However, these methods treat experiences of employee as same similar, while other experience may have different effects on the decision of an employee.

For estimating the duration of a person working in the company, this problem receives little attention from the community. Some naive approaches use the personal information of employees or previous working duration to predict the duration. These methods did not use the information from the job transition and cannot give the desired result.

In this paper, we also address the bottleneck problem that arises with the use of a fixed-length encoding vector. This is thought to become especially problematic for long and complex sequences, where the dimensionality of their representation would be forced to be the same as for shorter or simpler sequences. We improve the accuracy of predicting the next job title and working duration with attention mechanism to advantage weighting learned representations. Furthermore, we experience our method on real-world IT domain dataset and show the improvement over other state-of-the-art methods.

2. Problem Formulation

In this section, we formulate the problem by using mathematic equation and present the notations used throughout the paper.

Let $u \in U$ denote an employee, J_i^u denote user u's i-th working experience. For each experience, J_i^u=(Q_i^u, t_i^u) denote user u working as job title Q in t duration. User u experience sequence is denoted as vector $\overrightarrow{J^u}$ and can be summarized as:

$$\overrightarrow{J^u} = \{(J_1^u, J_2^u, ..., J_i^u) | u \in U\} \quad (1)$$

Connected with experiences, user u also have a list of skill S_k^u in their profile. We denote user u's skill set by:

$$S^u = \{(S_1^u, S_2^u, ..., S_k^u) | u \in U\} \quad (2)$$

We formulate the problem as follow:

Given: User u's working experiences sequence $\overrightarrow{J^u}$ and user u's skill set S^u. We have an input:

$$I^u = (\overrightarrow{J^u}, S^u) \quad (3)$$

Predict: User u's next job title Q_{i+1}^u and the duration t_{i+1}^u user u will working as job title Q_{i+1}^u. We have an output:

$$O^u = (Q_{i+1}^u, t_{i+1}^u) \quad (4)$$

Fig. 1 shows an example input and output of our model.

3. System overview

An overview of our model is given in Fig. 2. We detail our model by describing the following steps.

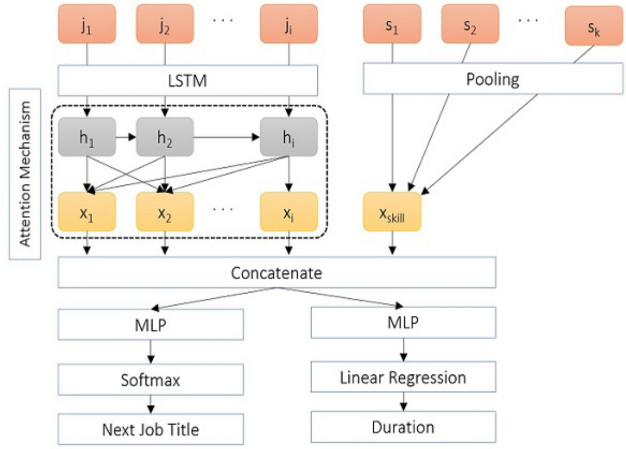

Fig. 2: The graphical representation of our model using Attention Mechanism in weighting the hidden states from LSTM layer to get latent information between job experiences

TABLE 1: Performance comparison of our model and other methods using mpr on our dataset

Data	Model	MPR	RSME
Job Experiences	Random Forest	0.152	0.914
	Logistic Regression	0.223	0.962
	LSTM	0.094	0.731
	NEMO	0.096	0.715
	LSTM-Attention	0.091	0.693
	NEMO-Attention	0.094	0.672
	Ours	**0.079**	**0.461**
Job Experiences and Skill	Random Forest	0.144	0.924
	Logistic Regression	0.192	0.913
	LSTM	0.092	0.671
	NEMO	0.085	0.624
	LSTM-Attention	0.089	0.586
	NEMO-Attention	0.082	0.512
	Ours	**0.071**	**0.347**

3.1. Encoding the context representation

To handle the rich forms of data (free text, numerical and categorical features), the data is preprocessed with the following methods. For the free text feature, such as job title and skill, we used the embedding method to transform a word into a vector. Then we computed the mean value of the embedding for every dimension respectively, in this way we got a fixed length of the vector for the free text of varying lengths.

Where skill encoder and job encoder can be some unsupervised textual embedding approaches such as Word2Vec [10], Skill2Vec [11], Job2Vec [12]. In our model, FastText [13] model is used to represent the job and skill to ensure that we can handle out of vocabulary issues while inference the model on actual data. We want to learn a representation of the context information from the user experience sequence. We create two encoder models for jobs and skills to learn their representation. Let $s_1^u, s_2^u, ..., s_k^u$ be the set of user u's skill embedding and $q_1^u, q_2^u, ..., q_i^u$ be the set of user u's job embedding. The embedding of job and skill will be represented as:

$$s_k = skill\_encoder(S_k) \quad (5)$$

$$q_i = job\_encoder(Q_i) \tag{6}$$

After having the text representation of job experience we integrate with the duration $j_i = (q_i, t_i)$ as an input to the LSTM layer.

User has a different number of skills, from as few as one skill to more than 20 skills. To ensure that each user has embedded with the same dimension, we use a pooling method across skill embedding vectors. In particular, we perform average-pooling to get a single skill vector.

$$x^u_{skill} = avg(s^u_1, s^u_2, ..., s^u_k) \tag{7}$$

3.2. Experience sequence representation

After the embedding, we feed the experience sequence into LSTM layer to learn the hidden representation of the position-specific sequential features. We refer the output of LSTM of this layer as h_1, h_2, h_g and then attention-mechanism is integrated with these outputs to get the final representation for user experience sequence. Bahdanau attention mechanism [14] is used to improve the performance of the encoder-decoder model for machine translation. The idea behind the attention mechanism was to permit the decoder to utilize the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all of the encoded input vectors, with the most relevant vectors being attributed to the highest weights. The attention mechanism can be formulated as follows:

$$e_t = \tanh(W_a h_i + b_a) \tag{8}$$

$$a_t = softmax(e_t) \tag{9}$$

$$x_g = \sum_{i=1}^{g} a_t(h_i) \tag{10}$$

Where W_a, b_a are training parameters, h_i means the hidden states learned from the LSTM layer, and x_g is the output vector for g-th job representation.

The job experience feature x_g is concatenated with the skill feature x^u_{skill} and becomes the input of two multilayer perceptron (MLP) [15].

3.3. Prediction module

In the prediction module, the predict next job title problem is regarded as a classification task. We optimize the loss function for predicting the next job title. With α_u is the true job title and $\hat{\alpha}_u$ is the predicted job title of an employee, the job title loss can be formulated as follows:

$$L^u_{job\ title} = -\sum_{i=1}^{N} \alpha_u \cdot log(\hat{\alpha}_u) \tag{11}$$

For the second task of predicting the duration, we treat this problem as linear regression. With β_u is the true duration and $\hat{\beta}_u$ is the predicted duration of an employee. We formulated the duration loss as follows:

$$L^u_{duration} = (\beta_u - \hat{\beta}_u)^2 \tag{12}$$

By combining the loss function for predicting next job title (11) and job duration (12), we construct our final loss function as below:

$$Loss = -\sum_{u \in U} (\theta \cdot L^u_{job\ title} + (1-\theta) \cdot L^u_{duration}) \tag{13}$$

4. Experiments

To verify the proposed model, we create a series of experiments including the comparison with other state-of-the-art methods and data ablation. The details are shown in the following subsections.

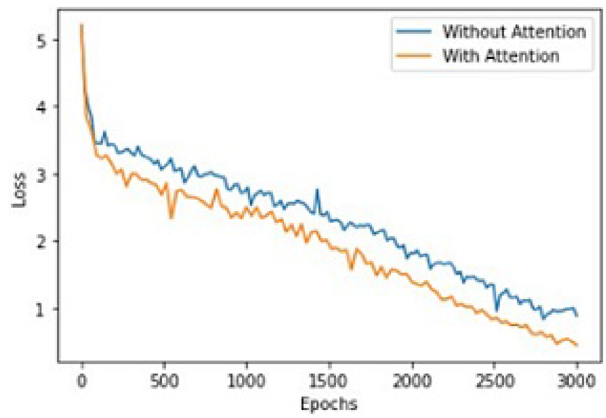

Fig. 3: Loss in training progress

4.1. Data Description

We need to build up a specific dataset for employee experience since the data is sensitive and no dataset is available. We use the real-world data from more than 10.000 employees in the IT domain, which were collected from a well-known online professional social platform. For preprocessing, were move employees with no experience or with more than 20 experiences reported in their profile. We also remove skills and job titles that appear less than 10 times in the dataset after formatting all the text.

4.2. Comparison with State-of-the-art methods

4.2.1. Baseline

Baseline models for job title prediction are non-sequential models (Logistic Regression, Random Forest) and deep learning models (LSTM, NEMO). We also tested modified versions of the deep learning model with attention mechanism (LSTM-Attention, NEMO-Attention) to compare the superiority of the attention mechanism in weighting the experiences. We experiment with our model with two different datasets. The first dataset uses job experiences as training features, the second dataset uses both job experiences and skills as training features. By using two datasets, we can illustrate the contribution of skill features in improving the overall result.

4.2.2. Metrics

Accuracy is a metric often used to evaluate classification models in machine learning. However, the accuracy metric does not distinguish between the likely job title and gives misleading results. To evaluate the prediction of next job title, Mean Percentile Ranking (MPR) [16] is used as our metric to take account of the rank of correct job title. For each data point, the model ranks all the job titles in order of likelihood, the high-rank job title being the most recommended job. Lets U_{test} is the set of members who have new position during the test period, we define the MPR metric as follow:

$$MPR(j) = \frac{1}{U_{test}} \sum_{u \in U_{test}} \frac{1}{|J|} rank(j_u^*) \tag{14}$$

Where $rank(j_u^*)$ is the rank of user u's actual job title in the ranking list. The ranking list is obtained by sorting the prediction scores of each job title from high possibility to low possibility. Lower MPR is more desirable because it indicates the model can rank the actual job title higher in the ranking list. In contrast, higher MPR indicates the actual job title has a low rank compared to another job title.

Root mean square error (RMSE) is widely used to measure the accuracy of a linear regression model. It can evaluate the difference between the actual duration and the predicted duration. The metric is formulated below:

$$RMSE = \sqrt{\sum_{u \in U_{test}} \frac{(\beta_u - \hat{\beta}_u)^2}{U_{test}}} \tag{15}$$

The performance of the baseline methods and our model is presented in Table 1. Our model outperforms baseline methods and achieves the highest result. A remarkable observation is that the overall performance increases significantly by applying attention mechanism to the traditional deep learning model (NEMO, LSTM).

Figure 3. shows the loss after 3000 training epochs of our model. By using the attention mechanism, the loss in the training progress drops faster and our model can achieve higher results.

5. Conclusion

In this paper, we conduct a comprehensive to explore the career path of employees at the individual level with attention mechanism, which improves the overall accuracy for both prediction tasks. The experiments confirm that, by using the attention mechanism, our method efficiently captures the intrinsic representation and profitably the connection of the employee's experience sequence better than baseline methods. Moreover, these representations are used to predict the next potential job title of an individual employee and predict the working duration at the next job title. Experimental results on real-world data show that, our model provides much better accuracy and is competitive with other state-of-the-art methods.

Acknowledgement

This research was supported by Orient Software and Neurond. We thank Orient Software and Neurond for providing us not only the best facilities but also insight and expertise that greatly assisted the research.

References

[1] S. T. Al-Otaibi and M. Ykhlef, "A survey of job recommender systems", *International Journal of Physical Sciences*, 7 (29), 2012, 5127–5142.

[2] Piad, Keno, Dumlao, Menchita, Ballera, Melvin, Ambat, Shaneth, "Predicting IT employability using data mining techniques", *Third International Conference on Digital Information Processing, Data Mining, and Wireless Communications*, 2016, 26-30.

[3] Corné de Ruijt et al, "Job Recommender Systems: A Review", *CoRR abs/2111.13576*, 2021. https://arxiv.org/abs/2111.13576.

[4] S. Choudhary, S. Koul, S. Mishra, A. Thakur, R. Jain. "Collaborative job prediction based on Naïve Bayes Classifier using python platform", *Collaborative job prediction based on Naïve Bayes Classifier using python platform*, 2016, 165-178.

[5] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme. "Factorizing personalized markov chains for next-basket recommendation". *Proceedings of the 19th International Conference on World Wide Web*, 2010, 811–820.

[6] Van Huynh, T., Van Nguyen, K., Nguyen, N. L.T., Nguyen, A. G.T. "Job Prediction: From Deep Neural Network Models to Applications", *RIVF Interna-tional Conference on Computing and Communication Technologies*, 2020, 1-6.

[7] Dr. Padmaja Pulicherla et al, "Job Shifting Prediction and Analysis Using Machine Learning", *Journal of Physics Conference Series*, 2019, 1-11.

[8] L. Li et al., "Nemo: Next career move prediction with contextual embedding", *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, 505-513.

[9] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong, "A hierarchical career-path-aware neural network for job mobility prediction", *The 25th ACM SIGKDD International Conference*, 2019, 14-24.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", *CoRR abs/1301.3781*, 2013. https://arxiv.org/pdf/1301.3781.pdf.

[11] D. Zhang et al. "Job2vec: Job title benchmarking with collective multi-view representation learning", *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019, 2763–2771.

[12] Le Van-Duyet, Vo Minh Quan, Dang Quang An, "Skill2vec: Machine Learning Approach for Determining the Relevant Skills from Job Description", *Computing Research Repository*, 2017, 1-3.

[13] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, T. Mikolov, "FastText.zip: Compressing text classification models", *CoRR abs/1612.03651*, 2016. https://arxiv.org/abs/1612.03651.

[14] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, *CoRR, abs/1409.0473*, 2014. https://arxiv.org/abs/1409.0473.

[15] M. W. Gardner, S. R. Dorling, "Artificial neural networks (the multilayer perceptron) A review of applications in the atmospheric sciences", *Atmospheric Environment*, 32, 1998, 2627-2636.

[16] Y. Hu, Y. Koren, and C. Volinsky. "Collaborative filtering for implicit feedback datasets", *Proceedings of the 8th IEEE International Conference on Data Mining*, 2008, 263–272.

Truong Dinh Thien is pursuing his Master at the University of Da Nang - University of Science and Technology, Vietnam. He has done bachelors in Software Engineering from Programme de Formation d'Ingénieursd' Excellence au Vietnam (PFIEV). He is working for the Research and Development department of Orient Software and Neurond. His research focuses on Machine Learning, Artificial Intelligence..

Ton Quang Hoang Nguyen has done bachelors in Embedded System Engineering from the University of Da Nang - University of Science and Technology, Vietnam. He is working for the Research and Development department of Orient Software and Neurond. He is the author of some scientific publications in international journals and conferences. His research focuses on Machine Learning, Computer Vision, and Artificial Intelligence.