ECAI 2023
K. Gal et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230319

What Wikipedia Misses About *Yuriko Nakamura*? Predicting Missing Biography Content by Learning Latent Life Patterns

Yijun Duan[a;*], **Xin Liu**[a], **Adam Jatowt**[b], **Chenyi Zhuang**[a], **Hai-Tao Yu**[d], **Steven Lynden**[a], **Kyoung-Sook Kim**[a] **and Akiyoshi Matono**[a]

[a]National Institute of Advanced Industrial Science and Technology (AIST), AIRC, Japan
[b]University of Innsbruck, Austria
[d]University of Tsukuba, Japan

Abstract. *Action-related KnowledGe* (AKG) is important for facilitating deeper understanding of people's life patterns, objectives and motivations. In this study, we present a novel framework for *automatically predicting missing human biography records in Wikipedia* by generating such knowledge. The generation method, which is based on a neural network matrix factorization model, is capable of encoding action semantics from diverse perspectives and discovering latent inter-action relations. By correctly predicting missing information and correcting errors, our work can effectively improve the quality of data about the behavioral records of historical figures in the knowledge base (e.g., biographies in Wikipedia), thus contributing to the understanding and study of human actions by the general public on the one hand, and can be considered as a new paradigm for managing action-related knowledge in digital libraries on the other. Extensive experiments demonstrate that the AKG we generate can capture well missing or "forgotten" human biography related information in Wikipedia.

1 Introduction

The study of human actions has long attracted the interest of scientists from different fields. For example, social scientists have been interested in the structure of life events and the role of individual agency and larger social forces in shaping individual life experiences [18, 12, 26]. Researchers in the information retrieval domain have found that many people use search engines not only to obtain information, but also to perform certain actions and achieve certain life goals [4]. Web knowledge bases, e.g., Wikipedia, contain such knowledge about the histories of people. In Wikipedia, many person entries contain the biography sections, making it an important data source for studying human behavior patterns [2, 9, 26]. However, the biographies in Wikipedia follow a long-tail distribution: the life of a small number of historical or contemporary figures is well documented and analyzed, while the life accounts of a vast majority of person entries are very limited. This makes current content in Wikipedia biased towards important individuals whose lives are well described; life stages such as those of *Yuriko Nakamura*[1] are essentially missing. Furthermore, it has been also recently found that among 1.5 million biographies in Wikipedia, only less than 19% are about women [24]. Closing this "gender gap" in Wikipedia requires much effort and volunteer activities. Clearly, some content could be translated from other Wikipedia versions (e.g., the Japanese Wikipedia) to alleviate this issue but often that content is equally limited, especially for smaller Wikipedia language editions. Therefore, in this study we aim to *uncover the unrecorded but highly probable biographical accounts of the "silent majority"*[2] *in Wikipedia.*

Our objective is then to understand, represent, and reason about actions to be able to calculate the probability that a particular person *u* performs action *a* at time *t*, and thus predict the missing yet highly plausible action records in Wikipedia (see Sec. 5.1). By this the content generation and editing activities in Wikipedia can be supported through provided hints and suggestions.

The automatic construction of Action-related KnowledGe (AKG), i.e., learning vector representations for actions and their relations from Wikipedia biographies is an important, challenging, and still *unexplored* problem. In this study, we first propose two matrix factorization models for AKG generation from different perspectives: *temporal context*[3] and *subject of action*. Based on these perspectives we make the following intuitive assumptions: (1) **H1**: actions that tend to occur at different life stages (i.e., have different temporal context, such as *entering high school* and *divorce*) are more likely to have different vector representations; (2) **H2**: actions performed by similar subjects (*e.g.*, *groups of scientists*) tend to have similar vector representations; (3) **H3**: an action's functionality[4] can be reflected by other actions that are complementary to it. We assume two (or more) actions are complementary to each other if they together satisfy a more general need of the subject (*e.g.*, *searching for a university* and *taking the entrance exam* are complementary actions for they together satisfy the subject's need of *receiving higher*

[*] Corresponding Author. Email: yijun.duan@aist.go.jp
[**] Dr. Chenyi Zhuang contributed to this work during his career in AIST, while he has retired at the date of publication.

[1] A Japanese jazz pianist who does not have her biography recorded in the English Wikipedia (https://en.wikipedia.org/wiki/Yuriko_Nakamura).
[2] We use this phrase to emphasize the long-tail distribution of person-related articles when ordered by the size of content in their biographies.
[3] Temporal contextual information about an action includes a description of the temporal period during which it is performed.
[4] We interpret the functionality of an action by the need which a subject wants to satisfy by performing such action.

education). Therefore, actions that are functionally complementary to similar actions are assumed to have similar vector representations. Our proposed two matrix factorization models (explained in Sec. 3) employ the above assumptions **H1** and **H2** respectively, and each generated AKG carries a single type of information.

Further, we propose a hybrid model based on neural network matrix factorization (NNMF) that considers multi-source signals and all above assumptions **H1**, **H2** and **H3** simultaneously (introduced in Sec. 4). This ensemble model not only utilizes the training data from all the perspectives mentioned above (subject, temporal context and functionality), but technically, it enjoys the following advantages: (1) allowing for unsupervised inter-action relation extraction in a label-free style (thus avoiding heavy data annotation process); (2) the matrix factorization module facilitates the learning process by injecting context knowledge into the neural network module (thus saving computation resources); (3) having capability of finding the optimal learning objective function rather than assuming it to be fixed.

We assume a complete action record in a Wikipedia biography is in the format of $\langle a, u, t\rangle$, which means person u performed action a at time t. Hereby, in the experiments we organized the task on missing full action records prediction in Wikipedia, *i.e.*, predicting whether the unrecorded full $\langle a, u, t\rangle$ triple truly occurs or not. Encouraging experimental results demonstrate that the AKG generated by the NNMF ensemble model can achieve the best prediction accuracy. In summary, the main contributions of our paper are as follows:

- To the best of our knowledge, this study is the *first* work devoted to resolving the incompleteness of Wikipedia biographies. We achieve this by using AKG automatically generated from web biographies. Our work will not only contribute to provide more complete data for the study of human behavior, history and sociology, but can also offer a new paradigm for managing action data in digital libraries.
- A novel neural network matrix factorization framework for AKG generation is proposed. The model is experimentally demonstrated to have strong prediction accuracy of missing biography data.

2 Problem Formulation of AKG Generation

Input: The set of biographies D crawled from Wikipedia. Let $U = \{u_1, u_2, ..., u_N\}$ denote the set of recorded people in D. For person u_i, the corresponding biography is defined as a chain of actions in his/her life $B(u_i) = \{(a_1, t_1) \rightarrow (a_2, t_2) \rightarrow ... \rightarrow (a_n, t_n)\}$, where t_i denotes the time of action a_i, and $t_1 < t_2 < ... < t_n$. In addition, t_i can represent both the *absolute* time (*e.g.*, A. D. 1968) or *relative* time (*e.g.*, 24 years old). $D = \{B(u_1), B(u_2), ..., B(u_N)\}$. Tab. 1 shows an example of actions from the biography of *Audrey Hepburn*, a 20^{th} century actress.

Output: Generated action-related knowledge $G = (A, R)$ where A is the set of all actions, R is the set of all inter-action relations. All elements in A and R are represented by dense vectors.

3 Single-view-based Action-related Knowledge Generation

3.1 Temporal Context-based Action-related Knowledge

In this section, we describe how we learn the vector representations of actions that carry temporal information. Based on the aforementioned hypothesis **H1** (see Sec. 1), actions that occur at similar times should have similar vectors. First, we construct the *action-time matrix* $G \in \mathbb{R}^{|A|\times|S|}$, where the value of entry G_{ij} in row i, and column j is the number of times that action i has been performed in time unit j. Here, $|A|$ and $|S|$ are the number of actions and time units, respectively. j can represent either an absolute date (*e.g.*, A. D. 2008) or a relative time (*e.g.*, 25 years old). We denote the vector of action i as a_i and the vector of time unit j as s_j; by the principle of non-negative matrix factorization [16], there should be $G_{ij} \approx a_i \cdot s_j$, $a_i \geq 0, s_j \geq 0$.

Furthermore, previous work suggests that people tend to exhibit a high degree of continuity and consistency in their actions they perform in neighboring time units [9]. Based on this assumption, the time vectors should satisfy $s_i \approx s_{i+1}$, where s_i and s_{i+1} denote the vectors of the adjacent time units i and $i+1$, respectively. Then, we have the following matrix factorization formula:

$$\min_{A_t, S\geq 0} ||G - A_t \cdot S^T||^2 + \alpha \cdot \sum_{i=1}^{|S|-1} ||s_i - s_{i+1}||^2 + \lambda \cdot (||A_t||^2 + ||S||^2) \quad (1)$$

Here, $A_t \in \mathbb{R}^{|A|\times d}$ is the learned action vector matrix, and $S \in \mathbb{R}^{|S|\times d}$ is the time vector matrix. Both action vector and time vector are d-dimensional. $||\cdot||$ represents the Frobenius norm. λ is the parameter that controls the norms of A_t and S to prevent overfitting. The time-based method constructs $G(A, R)$ as $G(A = A_t, R = \varnothing)$.

3.2 Subject-based Action-related Knowledge

In this section, we describe our method for learning action representations from the subject perspective, i.e., the perspective of the action performer. Hypothesis **H2** (see Sec. 1) tells us actions performed by similar subjects should have similar representations. We use two data sources: *subject-action matrix* and *link* information among person pages in Wikipedia. Let the subject-action matrix be denoted as $H \in \mathbb{R}^{|A|\times|U|}$, where each item H_{ij} denotes the number of times action i is performed by person j. We denote the set of action vectors as $A = \{a_1, a_2, ..., a_{|A|}\}$, and the set of subject vectors as $U = \{u_1, u_2, ..., u_{|U|}\}$. Similarly, there should be $H_{ij} \approx a_i \cdot u_j$, and $a_i \geq 0, u_j \geq 0$.

Naturally, the action of a certain person is often influenced by others, and different people have different influences. Highly influential people are more likely to have more detailed and comprehensive life records in Wikipedia. The previous studies [13] indicated that different persons in Wikipedia have diverse historical importance [5]. In addition, people who are perceived to be of high importance tend to have Wikipedia pages connected to other important Wikipedia articles [9]. These findings imply that subjects during the matrix factorization process should be paid different weights. Given the link structure in Wikipedia biographies, we use the PageRank algorithm [21] to calculate the importance $I(u_j)$ for subject u_j. Given the subject importance, we propose the following objective function:

$$\min_{A_u, U\geq 0} \sum_{i=1}^{|A|} \sum_{j=1}^{|U|} I(u_j)(H_{ij} - a_i^T u_j)^2 + \lambda \cdot (||A_u||^2 + ||U||^2) \quad (2)$$

Here, $A_u \in \mathbb{R}^{|A|\times d}$ and $U \in \mathbb{R}^{|U|\times d}$ are the learned action vector matrix and the subject vector matrix, respectively. $a_i = A_u[i]$, $u_j = U[j]$. The subject-based method generates $G(A, R)$ as $G(A = A_u, R = \varnothing)$.

[5] For example, based on [13], two examples of people with high historical importance are Napoleon and Albert Einstein.

Action sentence	Date	Age
After the Germans invaded the Netherlands in 1940, Hepburn adopted the pseudonym Edda van Heemstra, because an "English sounding" name was considered dangerous during the German occupation.	1940	11
By 1944, Hepburn had become a proficient ballet dancer.	1944	15
For her role in Roman Holiday, Hepburn was also the first actress to win an Oscar, a Golden Globe and a BAFTA Award for a single performance in 1954.	1954	25
In August 1988 , Hepburn went to Turkey on the immunization campaign.	1988	59
She was awarded the Presidential Medal of Freedom in recognition of her work as a UNICEF Goodwill Ambassador in late 1992.	1992	63

Table 1. A sample of 5 of the 62 action sentences (underlined words are data after pre-processing as input to our model) from Wikipedia biography of *Audrey Hepburn* (born 1929). **Date** and **Age** represent the absolute time and relative time, respectively.

4 Multi-view-based Action-related Knowledge Generation

If a subject performs an action a_1 and then performs action a_2, it is reasonable to infer that a_2 may satisfy the requirement that a_1 cannot satisfy. As the previously stated **H3** (see Sec. 1) says, a_1 and a_2 are assumed complementary to each other if they together satisfy a more general need of the subject. In this case, we judge that there exists a latent relation r between a_1 and a_2 as a reflection of such functional complementarity. On the other hand, r can also be understood in an intuitive way as an expression of the tendency for a_1 and a_2 to frequently co-occur in people's biographies. Naturally, the relation r can be defined as a tuple: $r \doteq \langle a_1, a_2 \rangle$, which means the vector representation of r is the concatenation of a_1's vector (head slot action vector) and a_2's vector (tail slot action vector). However, such definition restricts the relation to be represented by exactly two actions (*i.e.*, the 1-to-1 scenario) and cannot encode the relation among multiple actions (*i.e.*, the 1-to-N/N-to-1/N-to-N scenario).

To relax the above limitation, we now represent a relation r as a vector of dimension k, and define the relation matrix as $R_* \in \mathbb{R}^{m \times k}$, where m is the predefined number of relations. Then, given any two action vectors a_i and a_j, $I_{(a_i,a_j)} \in \mathbb{R}^{1 \times m}$ is defined as the indexing vector of the relation between a_i and a_j. All elements of $I_{(a_i,a_j)}$ take values in the range $[0, 1]$, which represents the fitness of the corresponding relation. In the extreme case, $I_{(a_i,a_j)}$ has only the l-th element close to one, whereas the remaining elements are close to zero, indicating that the relation between (a_i, a_j) can be approximated as the relation $R_*[l]$. Furthermore, the relation vector of (a_i, a_j) is computed as $r_{(a_i,a_j)} = I_{(a_i,a_j)} \cdot R_*$. Similarly, the relation vector of n actions $(a_1, a_2, ..., a_n)$ can be obtained by $r_{(a_1,...,a_n)} = I_{(a_1,...,a_n)} \cdot R_*$, and thus we can accommodate 1-to-N, N-to-1 and N-to-N scenarios.

For any action tuple (a_i, a_j), we assume that $I_{(a_i,a_j)}$ is influenced by all the time, subject, and text semantics related to a_i and a_j. We propose to use the Neural Tensor Network (NTN) [1] to learn how these factors affect, given its high effectiveness and efficiency in modeling input relations. Given the action-time matrix G and action-subject matrix H, we have

$$I(a_i, a_j) = \sigma(t_i^T W^{[1:m]} t_j + V[t_i||t_j] + b). \tag{3}$$

Here, $t_i = [G[i]; H[i]; i]$ and $t_j = [G[j]; H[j]; j]$, where $G[i]$, $H[i]$, and i denote the time, subject, and semantic information [6] of a_i, respectively. $W^{[1:m]}$ is a weight tensor, || denotes the concatenation operation, V is a weight matrix, b is a bias vector. $\sigma()$ is the Sigmoid activation function that maps the output of the neural tensor network to the range of element values in the relation indexing vector $I(a_i, a_j)$: [0,1].

Instead of using linear functions to fit the compatibility between an action tuple and a relation, we propose to adopt a two-layer feed-forward neural network FNN for such compatibility estimation. The parameters of this network can be learned in an entire data-driven manner from the probability transition matrix T, where T_{ij} denotes the probability of action a_j being performed after action a_i is performed and T can be easily constructed from the action trajectory set D (defined in Sec. 2). Our objective function is as follows:

$$Obj_{NN} \doteq \min_{A_*, R_* \geq 0} \sum_{(a_i,a_j)} (T_{ij} - f_{ij})^2, \tag{4}$$

Here, A_* is the action vector matrix and $a_i = A_*[i]$. In addition, we have f_{ij} = FNN($\vec{s_{ij}}$) and $\vec{s_{ij}} = [a_i; a_j; I(a_i, a_j)^T \cdot R_* \cdot Q]$. In particular, to increase the diversity of identified relations, we introduce the constant sparsity matrix $Q = (1 - \theta) \cdot eye(m) + \frac{\theta}{m} \cdot ones(m)$ [22, 29] in the factorization objective, where $eye(m)$ and $ones(m)$ are an $m \times m$ identity matrix, and an $m \times m$ matrix with all entries of 1s, respectively. θ is the parameter controlling the sparsity of Q [7]. The introduction of Q will lead to a smoother matrix factorization process, thus driving the resulting vectors in R_* to be diverse [22].

Further, we make use of prior contextual information (G, H) to enable the neural network to learn action representation from diverse perspectives. The final objective function is as follows:

$$\boxed{\begin{aligned} &\min_{A_*, R_*, S, U \geq 0} Obj_{NN} + \alpha \cdot (||G - A_* \cdot S||^2 + ||H - A_* \cdot U||^2) + \\ &\qquad \beta \cdot (||A_*||^2 + ||R_*||^2 + ||S||^2 + ||U||^2) \end{aligned}} \tag{5}$$

In our NNMF ensemble, the neural network module learns the optimal nonlinear objective function in a data-driven manner, whereas the matrix factorization module limits the solution space of the neural network by utilizing the prior context. In this way, the learned action vector contains multi-source context information and can encode relations among multiple actions. Finally, the generated knowledge $G(A, R)$ is denoted as $G(A = A_*, R = R_*)$.

[6] We first use the BERT model [8] to obtain the action semantics, then we cluster all action vectors using the mini-batch K-means algorithm to get the cluster id i. The number of clusters is set to be 500.

[7] As suggested in [22], $\theta = 0.8$.

5 Predicting Missing Biography Content by Generated AKG

5.1 Prediction Formulation based on a Generative Process

The prediction task T can be described as follows: Given a biography dataset X (X is the set of triples $\{u, t, a\}$, where $\{u_i, t_j, a_k\}$ indicates that person u_i performs action a_k at moment t_j), and the AKG: $G = (A, R)$ which is regarded as an external knowledge base constructed from an action-related corpus D; based on partial observable data X^o in X, the task is to predict the remaining non-observable data $X^{\neg o}$ in X. We model this prediction task using a generative process by placing Gaussian priors on latent feature vectors. The generation process of triple $\{u_i, t_j, a_k\}$ in X^o is simulated as follows: During the generation process, G provides auxiliary action information.

Algorithm 1 Generative process for predicting (u_i, t_j, a_k) in X^o

```
for each subject $\vec{u_i} \in U_X$ do
    draw latent subject vector $\vec{u_i} \sim N(0, \sigma_U^2 \mathbf{I})$
end for
for each time unit $\vec{t_j} \in T_X$ do
    draw latent time vector $\vec{t_j} \sim N(0, \sigma_T^2 \mathbf{I})$
end for
for each action $\vec{a_k} \in A_X$ do
    draw latent offset vector $\vec{o_k} \sim N(0, \sigma_A^2 \mathbf{I})$
    draw latent action vector $\vec{a_k} = \vec{o_k} + \vec{a_k^\triangle}$    ▷ $\vec{a_k^\triangle}$ is auxiliary action information from generated AKG
end for
for each subject-time-action pair $(\vec{u_i}, \vec{t_j}, \vec{a_k}) \in X^o$ do
    draw the *plausibility score* $X_{i,j,k} \sim N(\vec{u_i}^T \vec{t_j} + \vec{u_i}^T \vec{a_k} + \vec{a_k}^T \vec{t_j}, \sigma_X^2)$
end for
```

In Algo. 1, U_X, T_X, A_X are the set of subjects, times, and actions contained in X^o, respectively. $N(u, \sigma^2)$ denotes the Gaussian priors. $\vec{a_k^\triangle}$ is the auxiliary action information derived from G. Based on the proposed models, $\vec{a_k^\triangle}$ can be A_t (model 3.1), A_u (model 3.2) or A_* (model 4).

5.2 Full Triple Prediction

After the optimal parameters $u_i \in U_X$, $t_j \in T_X$, and $a_k \in A_X$ in Algo. 1 are learned, we can conduct predictions of unobservable data $X^{\neg o}$ in X. We first focus on the following prediction task: predicting the most likely existing full action triples (u_i, t_j, a_k) in $X^{\neg o}$. Under our aforementioned generative process (see Algo. 1), the possibility score P_{ijk} for triple (u_i, t_j, a_k) is computed as follows:

$$P_{ijk} = \vec{u_i}^T \vec{t_j} + \vec{u_i}^T \vec{a_k} + \vec{a_k}^T \vec{t_j} \tag{6}$$

For each missing possible combination (u_i, t_j, a_k) in $X^{\neg o}$, we compute its score and rank all triples based on such a score from high to low. The existence of top-ranked triples will be validated on ground truth data, and then the prediction effectiveness can be evaluated based on the metric ($Hits@k$), as introduced later.

6 Experiments

6.1 Dataset

We used an open large-scale Wikipedia biography dataset released by [2][8]. This dataset was published on January 2, 2014 and is a dump of English-language Wikipedia entries containing 242,970 biographies of people born after 1800. Each biography contains at least five actions, for a total of 2,313,867 actions. In addition, the data structure for each event is $\{person\_id, event\_id, year, age, terms, original\_sentence\}$, where $terms$ consists of meaningful words extracted from $original\_sentence$ and is used to compute the action's BERT vector [8] during pre-processing [9].

6.2 Data Preparation

We used 80% of the Wikipedia biography dataset [2] as the dataset D to learn AKG and the remaining 20% as the dataset X for the missing action prediction task (see Sec. 5.1). For X, we test for sparse training (existing biography for training / pseudo-missing biography for testing: 50% of X / 50% of X) and dense training settings (training / testing: 80% of X / 20% of X) separately. In addition, we test the prediction performance of the analyzed models in absolute time and relative time separately, because they are two fairly different temporal measures. Thus, we have four experimental settings in total: (sparse, absolute time), (dense, absolute time), (sparse, relative time), and (dense, relative time). For each setting, we used 5-fold cross validation to obtain the average model performance.

6.3 Analyzed Methods

First, the AKG generation model we propose consists of:

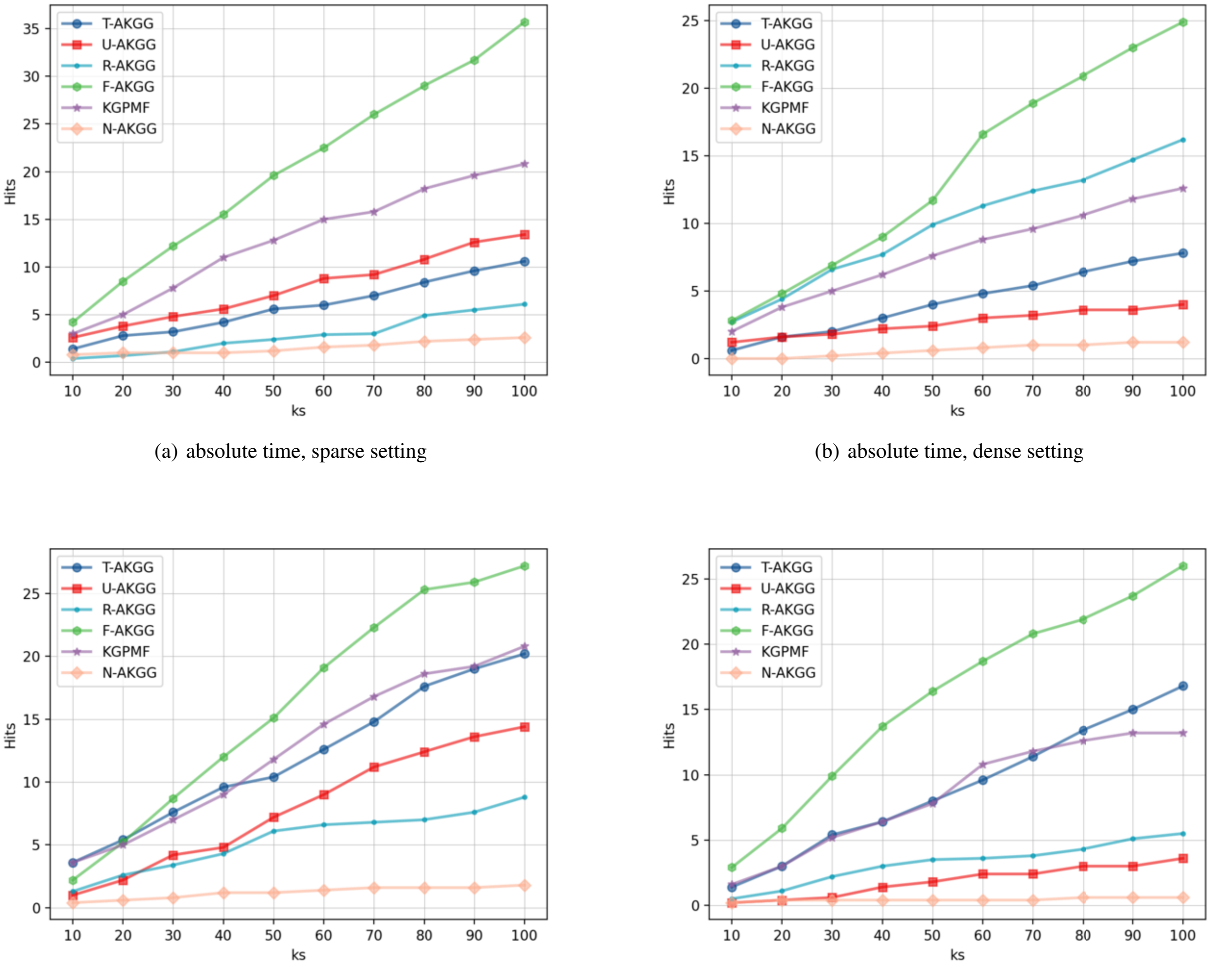
- **T-AKGG** (Time-based AKG generation) (see Sec. 3.1);
- **U-AKGG** (Subject-based AKG generation) (see Sec. 3.2);
- **F-AKGG** (Full AKG generation) (see Sec. 4).

In addition, we test the performance of the following baselines:

- **N-AKGG** (Non-action-related AKG generation). We do not use any information from the generated knowledge in the task of missing action prediction. This method was tested to verify the validity of AKG in prediction.
- **R-AKGG** (Functionality-based AKG generation [29]) A matrix factorization model that relies on mining inter-action relations described as *selective preference* [25]. However, **R-AKGG** restricts the relation to be represented by exactly two action vectors and cannot encode the relation among multiple actions. In addition, it assumes that the compatibility between action and relation is computed linearly.
- **KGPMF** (KG-based probabilistic matrix factorization [29]) The current state-of-the-art method for AKG generation, based on a probabilistic matrix factorization model. Different from **KGPMF**, we use different objective functions to implement different assumptions, and the neural network module in **F-AKGG** has a novel architecture.

[8] The dataset can be found at http://www.cs.cmu.edu/~ark/bio/

[9] The BERT model we used is a pre-trained model trained on Wikipedia, as published by Hugging Face [27].



(a) absolute time, sparse setting

(b) absolute time, dense setting

(c) relative time, sparse setting

(d) relative time, dense setting

Figure 1. Performance of all analyzed models for full action prediction in terms of $Hits@k$.

6.4 *Parameter Learning*

For Eq. (1), Eq. (2) and Eq. (5), *stochastic gradient descent* was used to approach the optimal parameters, and *multiplicative update rules* [17] were adopted to ensure that all parameters are non-negative. To minimize Eq. (5), we alternate between updating neural network parameters while fixing the AKG vectors, and updating the AKG vectors when fixing the network parameters. To learn the parameters in task T, we first convert the generative process of T into an equivalent probabilistic matrix factorization formula [20], and then solve it using the same strategy above.

6.5 *Evaluation Metrics*

In this study, we use $Hits@k(k = 10, 20, ..., 100)$ as our main evaluation criterion, where the meaning of $Hits@k$ is equivalent to $TPs@k$ (true positives@k). In our case, if a record of subject u conducting action a at time t does not appear in the test set, there exists the possibility that *this record is not included in Wikipedia*, rather than meaning that u has not conducted a in actual life. Thus, FPs (false positives) are not reliable in our scenario, which makes it difficult to report $Precision@k$ and $F - score@k$. Since the value of $Hits@k + FNs@k$ (FNs: False negatives) for each k is the same for all analyzed methods, reporting $Hits@k$ is equivalent to reporting $Recall@k = Hits@k/(Hits@k + FNs@k)$.

6.6 *Experimental Settings*

Based on a grid search (range: 0-100, step size: 10), the number of relations was set to 50, and the dimension of the relation vector was made equal to 20. The dimension of the action/time/subject vector was set to 50 for all methods. Inspired by [28, 29], in methods **U-AKGG**, **F-AKGG**, and **KGPMF**, we first cluster similar subjects into groups (group number: 100), and use group-level subject importance instead of individual importance to increase prediction accuracy. Other parameters are: $learning\_rate = 0.01$ and $\lambda = 0.001$. For the ensemble model **F-AKGG**, our 5-fold cross-validation suggests $learning\_rate = \lambda = 0.005$, $number\_of\_epochs = 2000$, and the hidden layer dimension is 20. For **T-AKGG**, $\alpha = 0.01$. For **R-AKGG** and **F-AKGG**, $\theta = 0.8$ (as suggested in [22]).

Action + Date	Most probable terms in action
music study (0s, 10s)	study, studied, conservatory, composition, school, piano, graduated
band formation (20s)	band, joined, began, formed, jazz, drummer, group, music
music performing (20s, 30s)	concert, held, live, music, song, performed, tour, show
album recording (30s)	jazz, recorded, recording, album, band, blues, played, records
album releasing (30s, 40s)	released, number, announced, album, release, single, hit, new album.
media interview (40s)	interview, music, said, work, wrote, critic, film, magazine
award receiving (50s)	award, album, best, year, music, nominated, records, received

Table 2. The corresponding descriptive terms of predicted actions. Due to space limitation for each action we only show up to 8 descriptive terms.

6.7 *Experimental Results (Full Triple Prediction)*

Fig. 1 illustrates the performance of all methods in the full triplet prediction task. Taken together, our NNMF ensemble model **F-AKGG** is observed to achieve the best prediction accuracy. We perform the following detailed analysis:

- Missing biography content prediction is a difficult and challenging task. The performance curve of **N-AKGG** shows that the average $Hits@100$ value of **N-AKGG** is only 2.2 and 0.9 in sparse and dense cases, respectively. This observation also hints at the sparseness of biographical data in Wikipedia.
- The prediction accuracy can be greatly improved when using AKG as a supplementary knowledge. This finding demonstrates that the study of AKG is useful in helping machines understand and encode human action. In particular, **T-AKGG**, **U-AKGG**, and **R-AKGG** are on average 6.5 times, 5.7 times, 3.0 times (sparse), and 12.3 times, 4.2 times, and 12.2 times (dense) more effective than **N-AKGG**, respectively. Thus, we conclude that *time*, *subject*, and *functionality* are all valid signals for AKG generation.
- With more training data, **R-AKGG** performs significantly better. The average $Hits@k$ value of **R-AKGG** in the dense setting increased by 57.7% compared with the sparse setting. When set to (dense, relative time), **R-AKGG** is generally the best performer among all baselines, implying that capturing the latent relations between actions requires large training data.
- In most cases, *time* signal is more effective and robust than *subject* signal. Overall, the average $Hits@k$ value of **T-AKGG** was 53.3% higher than that of **U-AKGG**. A very likely reason for this is that the temporal distribution of action in the training and test sets is much closer, whereas the difference between the subject distribution is more significant.
- The two methods that fuse all types of signals **F-AKGG** and **KGPMF** perform the best. Moreover, our proposed **F-AKGG** exhibits the best performance in almost every setting. This demonstrates the strong effectiveness of our NNMF construction framework in the full triplet prediction task.

7 Case Studies: Generated Partial Biography for *Yuriko Nakamura*

In this section we present predictions made by our generated AKG about the life of *Yuriko Nakamura*, who at the time of writing (April 2023) does not have a biographical entry in the English Wikipedia. This is a double partial triple prediction task: given subject u, predict action a and occurrence time t. Since *Yuriko Nakamura* was not included in our dataset, we set her vector representation as the average of the vectors of three similar Japanese female musicians[10] who we manually select from our dataset. Tab. 2 shows the top-10 predicted actions in the life of *Yuriko Nakamura* in chronological order predicted by **F-AKGG**. Note that to increase prediction accuracy we use a relatively wide temporal granularity: 10 years. The predicted actions were validated using her official homepage[11] and confirmed to have actually occurred based on her personal profile.

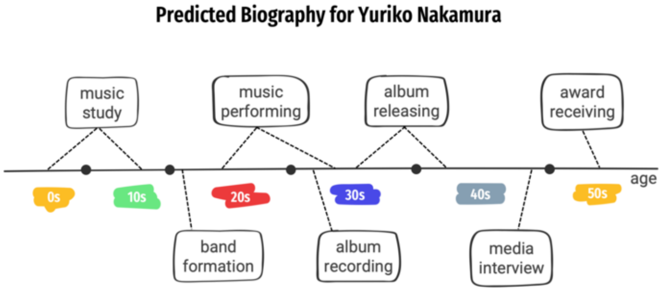


Figure 2. Top-10 predicted actions in the life of *Yuriko Nakamura* made by our generated AKG in chronological order. Each action is attached with its most probable occurrence time (e.g., 20s).

8 Related Works

This work lies in the field of biography mining. Related researches include the studies on the interplay between life events [18], the characteristics of life events that are highly transformative and iconic [12], the text generation from structured data with application to the biography domain [15], and the systematic differences in life structure across groups [26], etc. In particular, Wikipedia has been used extensively, for the task of disambiguation of named entities [6], the recognition of biographical sentences [5], the identification of latent biographical structure [2] and the summarization of typical life trajectories and events [9, 10], etc. This work is also related to two tasks: action mining and query-based actionable knowledge graph construction. The former task was designed to search for related actions associated with the query entity, whereas the latter task was designed to generate relevant actions and entity types (*e.g.*, thing, event) for the given query. Related works regarding these tasks include [23, 14, 3, 19]. Technically, our study is close to the knowledge graph embedding (KGE) domain (especially, multi-view KGE). Due to space limitation, we cannot provide a complete review of all important works of KGE. More detailed and comprehensive reviews of KGE can be found in [11, 7].

[10] These musicians are *Akiko Yano* (https://en.wikipedia.org/wiki/Akiko_Yano), *Ringo Sheena* (https://en.wikipedia.org/wiki/Ringo_Sheena) and *Toshiko Akiyoshi* (https://en.wikipedia.org/wiki/Toshiko_Akiyoshi).

[11] https://yurikopia.com/disco/

9 Conclusions

The objective of enhancing our understanding of people's behavior patterns and life stages has attracted the interest of researchers in multiple disciplines. In this study, we make the first attempt to tackle the problem of predicting missing data of biographical corpus on Internet. By recovering biography records of the so-called "silent majority", we expect that the Wikipedia content writing and editing processes can be further supported. We also believe that the automatic story and timeline generation techniques can be enhanced when incorporating our method. Finally, researchers from sociology and history and other related fields should be able to conduct more generalized and less biased intellectual enquiry when having more complete data. From a technical side, we propose a novel multi-view neural network matrix factorization framework for recovering and in general understanding sequences of human actions that constitute biographical data. The high prediction accuracy of our approach has been demonstrated through extensive experiments and case studies.

Our research can not only assist social scientists and historians in the analysis of human actions and life decisions, but also contribute new technical innovations into the field of action relation extraction and Wikipedia editing support. In the future, we will focus on exploring ways to utilize the expressiveness of our approach in other sociological problems.

10 Acknowledgments

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). We would also like to acknowledge the support from JSPS Grant-in-Aid for Scientific Research (21K12042). Particularly, the corresponding author would like to express his deep gratitude to Dr. Chenyi Zhuang for his great contribution to this work.

References

[1] Yunsheng Bai, Hao Ding, Song Bian, Ting Chen, Yizhou Sun, and Wei Wang, 'Simgnn: A neural network approach to fast graph similarity computation', in *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 384–392, (2019).

[2] David Bamman and Noah A Smith, 'Unsupervised discovery of biographical structure from text', *TACL*, **2**, 363–376, (2014).

[3] Roi Blanco, Hideo Joho, Adam Jatowt, Haitao Yu, and Shuhei Yamamoto, 'Overview of ntcir-13 actionable knowledge graph (akg) task', in *Proceedings of the NTCIR-13 Conference*, (2017).

[4] Andrei Broder, 'A taxonomy of web search', in *ACM Sigir forum*, volume 36, pp. 3–10. ACM New York, NY, USA, (2002).

[5] Mike Conway, 'Mining a corpus of biographical texts using keywords', *Literary and Linguistic Computing*, **25**(1), 23–35, (2010).

[6] Silviu Cucerzan, 'Large-scale named entity disambiguation based on wikipedia data', in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 708–716, (2007).

[7] Yuanfei Dai, Shiping Wang, Neal N Xiong, and Wenzhong Guo, 'A survey on knowledge graph embedding: Approaches, applications and benchmarks', *Electronics*, **9**(5), 750, (2020).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, (2018).

[9] Yijun Duan, Adam Jatowt, and Katsumi Tanaka, 'Discovering typical histories of entities by multi-timeline summarization', in *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 105–114, (2017).

[10] Yijun Duan, Adam Jatowt, and Katsumi Tanaka, 'History-driven entity categorization', in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pp. 349–364. Springer, (2019).

[11] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, Sabrina Kirrane, Sebastian Neumaier, Axel Polleres, et al., 'Knowledge graphs', *arXiv preprint arXiv:2003.02320*, (2020).

[12] Dennis P Hogan and Nan Marie Astone, 'The transition to adulthood', *Annual review of sociology*, **12**(1), 109–130, (1986).

[13] Adam Jatowt, Daisuke Kawai, and Katsumi Tanaka, 'Predicting importance of historical persons using wikipedia', in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 1909–1912, (2016).

[14] Xin Kang, Yunong Wu, and Fuji Ren, 'Tua1 at the ntcir-13 actionable knowledge graph task: Sampling related actions from online searching'.

[15] Rémi Lebret, David Grangier, and Michael Auli, 'Neural text generation from structured data with application to the biography domain', *arXiv preprint arXiv:1603.07771*, (2016).

[16] Daniel D Lee and H Sebastian Seung, 'Learning the parts of objects by non-negative matrix factorization', *Nature*, **401**(6755), 788–791, (1999).

[17] Daniel D Lee and H Sebastian Seung, 'Algorithms for non-negative matrix factorization', in *Advances in neural information processing systems*, pp. 556–562, (2001).

[18] Lee A Lillard and Linda J Waite, 'A joint model of marital childbearing and marital disruption', *Demography*, **30**(4), 653–681, (1993).

[19] Xinshi Lin, Wai Lam, and Shubham Sharma, 'Cuis team for ntcir-13 akg task', in *Proceedings of the 13th NTCIR Conference*, (2017).

[20] Andriy Mnih and Russ R Salakhutdinov, 'Probabilistic matrix factorization', in *Advances in neural information processing systems*, pp. 1257–1264, (2008).

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, 'The pagerank citation ranking: Bringing order to the web.', Technical report, Stanford InfoLab, (1999).

[22] Alberto Pascual-Montano, Jose Maria Carazo, Kieko Kochi, Dietrich Lehmann, and Roberto D Pascual-Marqui, 'Nonsmooth nonnegative matrix factorization (nsnmf)', *IEEE transactions on pattern analysis and machine intelligence*, **28**(3), 403–415, (2006).

[23] Md Mostafizur Rahman and Atsuhiro Takasu, 'Tlab at the ntcir-13 akg task', in *Proceedings of the NTCIR-13 Conference*, (2017).

[24] Francesca Tripodi, 'Ms. categorized: Gender, notability, and inequality on wikipedia', *New Media & Society*, 14614448211023772, (2021).

[25] Tim Van de Cruys, 'A neural network approach to selectional preference acquisition', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 26–35, (2014).

[26] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier, 'It's a man's wikipedia? assessing gender inequality in an online encyclopedia', *arXiv preprint arXiv:1501.06307*, (2015).

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., 'Huggingface's transformers: State-of-the-art natural language processing', *ArXiv*, arXiv–1910, (2019).

[28] Yao Wu, Xudong Liu, Min Xie, Martin Ester, and Qing Yang, 'Cccf: Improving collaborative filtering via scalable user-item co-clustering', in *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 73–82, (2016).

[29] Chenyi Zhuang, Nicholas Jing Yuan, Ruihua Song, Xing Xie, and Qiang Ma, 'Understanding people lifestyles: Construction of urban movement knowledge graph from gps trajectory.', in *IJCAI*, pp. 3616–3623, (2017).