Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Paths of A Million People: Extracting Life Trajectories from Wikipedia

Ying Zhang*, Xiaofeng Li*, Zhaoyang Liu, Haipeng Zhang†

ShanghaiTech University
{zhangying12022, lixf2022, liuzhy2023, zhanghp}@shanghaitech.edu.cn

Abstract

The life trajectories of notable people have been studied to pinpoint the times and places of significant events such as birth, death, education, marriage, competition, work, speeches, scientific discoveries, artistic achievements, and battles. Understanding how these individuals interact with others provides valuable insights for broader research into human dynamics. However, the scarcity of trajectory data in terms of volume, density, and inter-person interactions, limits relevant studies from being comprehensive and interactive. We mine millions of biography pages from Wikipedia and tackle the generalization problem stemming from the variety and heterogeneity of the trajectory descriptions. Our ensemble model **COSMOS**, which combines the idea of semi-supervised learning and contrastive learning, achieves an F1 score of **85.95%**. For this task, we also create a hand-curated dataset, *WikiLifeTrajectory*, consisting of 8,852 *(person, time, location)* triplets as ground truth. Besides, we perform an empirical analysis on the trajectories of 8,272 historians to demonstrate the validity of the extracted results. To facilitate the research on trajectory extractions and help the analytical studies to construct grand narratives, we make our code, the million-level extracted trajectories, and the *WikiLifeTrajectory* dataset publicly available.

Introduction

Life trajectories (Elder Jr 1994) of notable individuals have profound and crucial implications. These people choose where to study, live, work, campaign, and spend the rest of their lives (Elder Jr 1994; Doherty 2009; Schich et al. 2014), while engaging in social interactions (Becker 1974). Throughout these processes, new ideas emerge and spread (Elder, Johnson, and Crosnoe 2003), communities are established, clusters are formed (Schatzki 2022), and technologies are created (Schatzki 2019). Schich et al. (2014) analyze the birth and death places of more than 150,000 notable people and reveal how cultural centers evolve over a span of more than 2,000 years. Studies also pay attention to scientists' relocation, and discover the shifts in scientists' career choices (Deville et al. 2014) and the power comparisons between nations (Verginer and Riccaboni 2020), and how these moves affect scientific productions and progresses (Zucker and Darby 2009). Additionally, the whereabouts of politicians imply political influences (Goldsmith, Horiuchi, and Matush 2021), alliances (Doherty 2009), and regional socioeconomic developments (Boianovsky 2018).

However, the trajectory data is scarce in terms of volume, density, and inter-person interactions, limiting previous studies from being comprehensive and interactive. For instance, the aforementioned cultural history study (Schich et al. 2014) does not go beyond birth and death places, while the places people choose to study and work in their prime, may better shape cultural centers. Besides, with these intermediate points, the life trajectories can be "densified" to tell more complete stories. Furthermore, if social interactions were available, the life trajectories would literally intersect, transforming into much richer dynamic spatio-temporal networks of people. Many other studies that rely on data of small scales (Morrissey 2015; Doherty 2009), can hardly be extended to grand narratives across time and space. Their data sources, such as hand-curate databases (Hautala and Jauhiainen 2019), government websites (Doherty 2009; Goldsmith, Horiuchi, and Matush 2021), and freebase (Schich et al. 2014), can hardly provide the desirable dataset, and this is the reason that we turn to the entire Wikipedia. According to our statistics, the raw Wikipedia dumps[1] contain 1,930,519 biography pages and on average each page has 11 locations and 15 time entities, serving as a convenient source of life trajectories.

It is challenging to fully utilize Wikipedia for the trajectory extraction task, due to the variety and heterogeneity of the trajectories. If we directly apply Named Entity Recognition (NER) tools to detect elements and construct *(person, time, location)* triplets, only 30% of the triplets are entirely correct, according to our estimation. To improve this, context information should be considered. Snippets (1) and (2) in Figure 1(a) showcase the importance of context. Snippet (2) highlights an incorrect extraction candidate – *(Bob Hayes, 1964, USA)*, while if we have the sports event context, we would know the location should be Tokyo as shown in snippet (1), instead of the USA which *Hayes* is representing. However, it is not easy to utilize the context information,

*These authors contributed equally.

†Corresponding author.

[1]We use the Wikipedia dumps of June 1, 2023, from https://dumps.wikimedia.org/.

since there is a large variety of contexts and different contexts may suggest different extraction patterns. We examine trajectories on 10 biography pages (out of the 1,930,519 ones in total) and the results already contain 37 categories. As shown in Figure 1(b), the contexts suggest a relatively even distribution, with various types covering 83%, apart from work and speech. Therefore, for any work that uses the contexts for such triplet extractions, the labeled training data would inevitably contain a very small portion of all possible context types. This would be problematic when classifying triplets with unfamiliar contexts, giving rise to an overfitting/generalization issue. This may explain the low Recall of the traditional rule-based method for a similar task (Lucchini, Tonelli, and Lepri 2019). In the study by Vempala and Blanco (2020), samples (also with contexts) were labeled from only 100 biography pages, focusing solely on the trajectories of the person who is the subject of the biography. The F1 score was 74% and if they test on samples from other people, the performance is expected to be lower. The generalizability may be improved, if the classifier knows what the contexts of triplets in the wild look like beyond the very limited labeled triplets, which rhymes with the idea of semi-supervised learning (Van Engelen and Hoos 2020).

Despite the heterogeneity of the trajectories suggested in Figure 1(b), similar contexts often imply similar extraction rules and vice versa. For instance, the contexts of snippets (1) and (3) in Figure 1(a) are similar (both about sport events), suggesting the same way of extraction, while the dissimilarity between the context of snippet (1) and that of snippet (4) (about birth and study) indicates the way of extraction for snippet (1) does not apply to snippet (4). Therefore, capturing the similarity and dissimilarity between training samples may help improve the classifier and contrastive learning shares this ideology. It creates pairs of similar examples and dissimilar examples and trains a model to distinguish between them (Khosla et al. 2020).

It is worth noting that besides Wikipedia, news articles are another popular source for similar extraction tasks (Piskorski et al. 2020; Peng et al. 2024). However, apart from the apparent differences such as shorter duration (mostly after the 17th century) and somehow biased coverage towards certain people such as politicians, artists and athletes (Gebhard and Hamborg 2020), news articles have their distinct characteristics. For a particular trip of a celebrity, there may be several news articles covering it and the key elements may be scattered in one article and among articles (Peng et al. 2024). For Wikipedia's biographies that we work on, one trip usually appears only on one page and its descriptions are often clustered in one place on that page. Therefore, a previous framework of extracting celebrity trips from news articles, CeleTrip (Peng et al. 2024), relying on capturing long-range cross-document dependency between key elements within one article and across articles with Word-Article graph, may not work well in our scenario. Instead, local semantics may be effective here.

In this paper, we propose **COSMOS** (**CO**ntrastive learning and **S**emi-supervised learning **MO**del for extracting **S**patio-temporal life trajectory) to accomplish the task of trajectory extraction from Wikipedia. Before it goes to work,

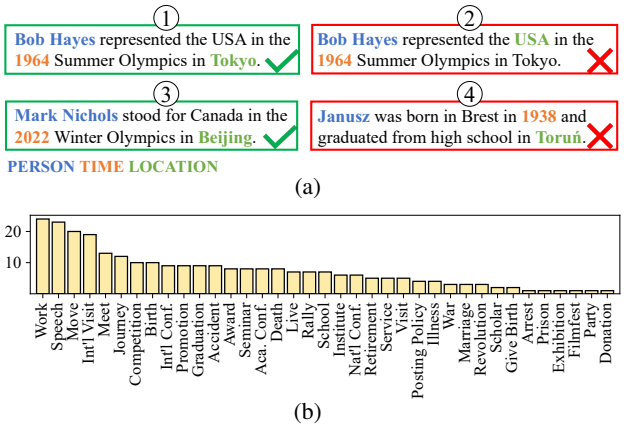


Figure 1: (a) Example of extracted triplets and their contexts, where green boxes represent correct trajectory information and red boxes represent incorrect ones. (b) The distribution of different trajectory types.

we extract the *(person, time, location)* triplets as candidates for it to classify. A correct triplet means that this *person* is actually in this *location* at this *time*. As mentioned above, in a person's biography, there could be m names, n times, and k locations and this gives us a maximum of mnk combinations as candidate triplets. To reduce the problem space, we take the heuristic to restrict the elements of a triplet to be in the same sentence and apply a syntax tree to further remove unreasonable combinations, as suggested by Liu, Hua, and Zhou (2020). According to our estimation, the remaining triplets would cover most of the correct ones. These triplets, combined with their corresponding contexts are then fed into **COSMOS**. COSMOS will classify these candidate triples into two categories ('trajectory' or 'not trajectory') based on the triples themselves and their context. The representation of a triplet and its context is obtained from BERT (Devlin et al. 2018). As discussed earlier, to mitigate overfitting and to exploit the dis/similarity among training samples, we combine semi-supervised learning (Van Engelen and Hoos 2020) and contrastive learning (Khosla et al. 2020), in a joint training framework.

Additionally, for this task, we create a first dataset annotating the general life trajectories, *WikiLifeTrajectory*, containing 8,852 triplets.

We summarize our contributions as follows.

- We formally formulate the task for life trajectory extraction from Wikipedia, and construct a curated dataset *WikiLifeTrajectory* dataset for this task. Though we perform our experiments on Wikipedia biographies, similar methods can be extended to other biography content.
- We design an effective method, **COSMOS**, which combines the idea of contrastive learning and semi-supervised learning to improve the generalization of the model facing diverse trajectory contexts. **COSMOS** extracts life trajectories and outperforms all baselines with an F1 score of 85.95% on the dataset.
- We make our framework, the million-level extracted

trajectories, and the *WikiLifeTrajectory* dataset publicly available[2]. The dense and interactive trajectories with wide coverage will support mobility analysis from case studies to large-scale modeling. Besides, we conduct an empirical analysis on the trajectories of historians to show the potential of our data.

Related Work

Life Trajectory Analysis

Life trajectory data has extensive applications in the field of social sciences (Yen et al. 2021). For instance, Schich et al. (2014) utilize the birth and death locations of more than 150,000 historically notable individuals to explore the evolution of culture centers. In addition to cultural history, trajectory information of specific populations is also used for analysis for various purposes. The trips of U.S. presidents are analyzed to understand presidential policy (Kernell 2006), while the migration patterns of artists can serve government decision-making (Hautala and Jauhiainen 2019). Life trajectories are also helpful in understanding human behavioral patterns. For example, Kleinepier, de Valk, and van Gaalen (2015) investigate the life trajectories of Polish immigrant families to identify factors influencing different family paths.

However, the current life trajectory data suffers from limitations in temporal scope, spatial coverage and inter-person interaction. These limitations emphasize the monotonous nature of the available data, which lacks diversity and comprehensiveness, restricting the ability of existing life trajectory data for large-scale, fine-grained, and networked analysis.

Spatio-Temporal Knowledge Extraction

There have been several studies trying to extract spatial or temporal knowledge related to people from different information sources. The most relevant to our task is the study working on text corpus. A relatively early study detects the type and timing of users' life events from social media based on hand-crafted features and traditional machine learning models (Dickinson et al. 2015). With the development of neural networks such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), researchers use deep learning models like Bi-LSTM to build life logs with timelines for individual users (Yen, Huang, and Chen 2019).

Recently, there are also some efforts dedicated to the extraction of spatio-tempoal knowledge. Lucchini, Tonelli, and Lepri (2019) extract births, deaths, and migrations of notable people on year-level. Their approach focuses on extracting knowledge from semantic roles defined by FrameNet, and only considers 29 frames "related to movements", resulting in a low Recall in general life trajectories (Lucchini, Tonelli, and Lepri 2019). Peng et al. (2024) detect celebrity trips from the news by modeling the news articles as graphs to capture the long-range cross-document dependency. However, trajectories in Wikipedia are more compact within single biography pages, which may make the sequential models more suitable than graph-based models for modeling text. In addition, Vempala and Blanco (2020) extract spatial timelines of the top 50 wealthiest individuals and top 50 multiple Olympic medalists from their biography pages based on LSTM. When applied to larger population groups and more diverse trajectory types, the model's generalization ability may be constrained if relying solely on limited labeled data. To mitigate this limitation, we introduce additional unlabeled samples during model training and guide the model to classify the samples through comparison.

[2]https://github.com/ZhangDataLab/COSMOS

Problem Statement

Before we formalize the core problem, we describe the heuristic which is briefly mentioned in the Introduction. Since we want to find where and when someone has appeared on a Wikipedia page, we first retrieve all possible combinations of *(person, time, location)* from the page as candidate triplets and then identify triplets representing someone's trajectory. To obtain these candidate triplets, we employ NER tools to recognize all entities related to person, time and location on one page, following by combining them to *(person, time, location)* triplets based on the structures of sentences, accompanied by the paragraphs they come from. The task now becomes: for each candidate triplet *(person, time, location)*, decide whether *person* has actually appeared in *location* at *time*, given the corresponding paragraph.

Given a triplet $t = (person, time, location)$ and its corresponding paragraph p, we train a model f with trainable parameters Θ to classify t into binary category $y \in \{0, 1\}$, with $y = 1$ indicating that $person$ has appeared in $location$ at $time$, and $y = 0$ being the opposite:

$$f : \{t, p, \Theta\} \rightarrow y. \tag{1}$$

Our pipeline for extracting candidate triplets will be described in Section Dataset.

Dataset

To the best of our knowledge, there is no dataset annotating the general life trajectories we are interested in, so we create our own ground truth dataset. This section describes how we construct the *WikiLifeTrajectory* dataset, which consists of two manually annotated datasets, namely "Regular" and "Representative".

Data Collection

The data is from the biography pages of the English Wikipedia. The list of people and the corresponding links to the Wiki pages are from Wikidata (Möller, Lehmann, and Usbeck 2022). In all, we have 1,930,519 people and their biography pages.

Data Processing and Labeling

This section describes how we extract and annotate candidate triplets in *(person, time, location)* format from biography pages.

Extracting Triplets We design a pipeline to extract candidate triplets from biography pages. The usability of this pipeline is also evaluated in real-world scenarios at the end of this section. Previous research utilizes NER tools to identify entities in sentences and employs a sliding window approach to select time and location entities related to a person (Vempala and Blanco 2020). However, this method ignores the structure of sentences and will introduce many meaningless connections between entities (Liu, Hua, and Zhou 2019). To address this limitation, we draw inspiration from Liu, Hua, and Zhou (2020) and construct parse trees to connect entities identified by NER. Ultimately, we obtain candidate triplets in the form of *(person, time, location)*.

In our extraction pipeline, we first use SpaCy[3] to identify entities. We retain sentences that only contain time and location entities as target sentences to ensure that we can connect these entities by parse trees. For each target sentence, we classify entities into four sets using SpaCy's classifications: *person* (personal pronoun and name entity PERSON), *time* (name entities DATE, TIME, and DURATION), *location* (name entities GPE, LOC, EVENT, FAC, and ORG), and *verb*. In order to find the logical connection between entities, we refer to the previous approach (Liu, Hua, and Zhou 2020). In each target sentence, we construct multiple (person, verb), (time, verb), and (location, verb) pairs based on the entities. We define a parse-tree-based distance metric to estimate the relevance of different pairs. The distance is calculated as the minimum path from each entity in a pair to their lowest common ancestor (LCA), and the pair with the minimum distance is considered the most relevant. Based on this comparison method, for each person entity in the *person* set, we identify the relevant (person, verb*) pair. According to verb*, we determine the most relevant time and location entities. Finally, the verb entity serves as a bridge connecting these three entities, giving us candidate triplets of the form *(person, time, location)*.

Finally, we compare the number of trajectories mentioned in the original pages with those extracted from the target sentences to measure the coverage of our method. Four biography pages with a total of 106 trajectory descriptions are manually checked, and our extraction pipeline can cover at least 85% (specifically, 86.11%, 90.00%, 85.71%, and 86.96% for each page, respectively) of the trajectories mentioned on different pages. The uncovered trajectories contain ambiguous expressions of time or location, such as "several years later", which inherently represent vague trajectories and are even hard to recognize by humans.

Annotating Triplets As mentioned in the Introduction, we collect trajectories from 10 biography pages and find that trajectories vary due to the variety of occupations and life stages. These trajectories are manually annotated in a *(person, time, location)* triplet format and labeled as positive ($y = 1$, defined in Problem Statement), resulting in the dataset "Regular". In addition, to cover more representative trajectory types when constructing the ground truth, we sample and annotate the triplets extracted by our extraction tool to obtain the dataset "Representative".

[3]https://spacy.io/

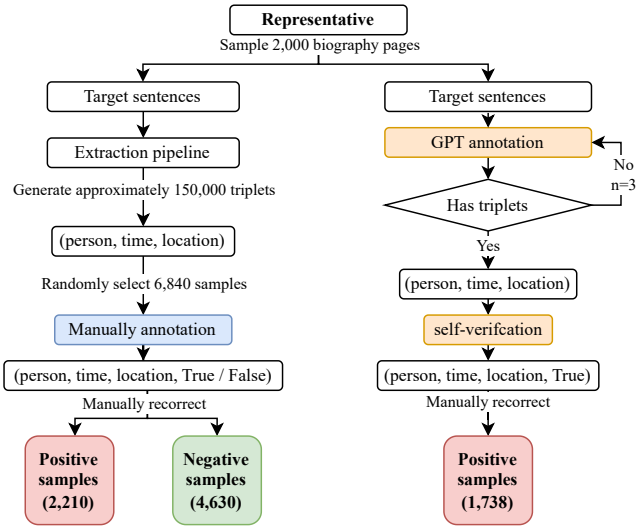


Figure 2: The flowchart illustrates the process of annotating the "Representative" dataset to obtain triplets and their corresponding labels.

Given that trajectory triplets' contexts are often associated with occupations (Li et al. 2017), we employ a stratified sampling and annotation strategy based on occupation to collect representative trajectories. We first identify the top 300 occupations based on their occurrence frequencies from people we collect, and then select additional 2,000 biography pages through stratified sampling based on the proportion of each occupation.

After extracting the candidate triplets through our extraction tool, we adopt a mixed annotation approach utilizing both human annotators and GPT-3.5, as recent studies utilizing GPT for data annotation have piqued our curiosity regarding the effectiveness of GPT-based data labeling (Thapa, Naseem, and Nasim 2023). From biography pages in "Representative", we choose individuals with longer page content and feed target sentences we extract to GPT-3.5 for annotation. Drawing inspiration from previous studies (Wang et al. 2023; Han et al. 2023), we design an instructive prompt[4] for GPT to generate trajectory triplets from the given target sentences following a self-verification mechanism. Though more than half of the trajectories will be lost after GPT's self-verification, the Precision of resulting annotations exceeds 90%.

Ultimately, we obtain 1,738 triplets labeled as positive ($y = 1$) annotated by GPT. Additionally, three undergraduate students with English proficiency annotate a total of 6,840 triplets, using target sentences distinct from those fed to GPT, where 2,210 triplets are labeled as positive and 4,630 triplets are labeled as negative ($y = 0$). A graduate student then verifies these results. The corresponding workflow is shown in Figure 2.

After the above annotation process, the "Representative" consists of 3,948 positive triplets and 4,630 negative triplets,

[4]See the Appendix for the prompts we provide for GPT.

while the "Regular" consists of 274 positive triplets. The triplets, along with the paragraphs and biography pages they originated from, are combined to form a complete JSON formatted dataset (i.e. the *WikiLifeTrajectory* dataset). Additionally, we randomly extract 50,000 unlabeled triplets from biography pages for subsequent semi-supervised learning in our method.

Method

The structure of COSMOS is shown in Figure 3. As described in Introduction and Problem Statement, for any triplet $t = (person, time, location)$, COSMOS learns its overall representations ($\mathbf{h}_{ce}$) through an ensemble model and then predicts the triplet label. When training the model, we introduce a combination of contrastive learning and semi-supervised learning. The following sections elaborate (1) the design of the ensemble model and how we integrate (2) contrastive learning and (3) semi-supervised learning into our model.

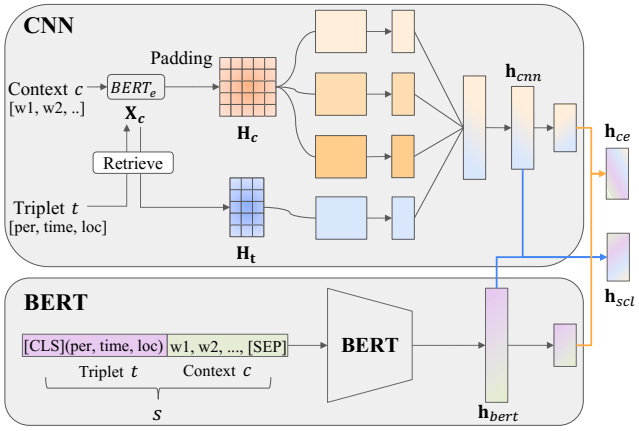


Figure 3: The framework of COSMOS. COSMOS learns the representations of triplets and their contexts through parallel CNN and BERT, and then classifies them based on the resulting representations.

Representation of Triplet

In life trajectory extraction scenarios, achieving a balance of Recall and Precision is crucial for accurate and reliable extraction. However, during our experimental analysis, we observe that individual models based on CNN or BERT exhibit contrasting performance characteristics (see Experimental Results). While the CNN model demonstrates a higher Precision, resulting in more precise predictions, the BERT model exhibits a higher Recall, capturing more relevant information. This discrepancy in performance motivates us to explore the potential benefits of integrating these two models into a hybrid approach.

Given a triplet t and its corresponding paragraph p, we will first match the sentences, which contain the target time or location in t, from p to compose the triplet's context c. Context c is specifically represented as a sequence of words. We then combine t with its context c using CNN and BERT in parallel (consisting of a joint training pipeline) to learn an overall representation of t.

Representation from CNN We first use the pre-trained English BERT model (denoted as $BERT_e$ in Figure 3) to extract word embeddings of c and obtain $\mathbf{X}_c \in \mathbb{R}^{n \times d}$, where n is the sequence length of c and d is the dimensions of embeddings. Then, we pad $\mathbf{X}_c$ to $\mathbf{H}_c \in \mathbb{R}^{k_1 \times d}$. Meanwhile, we retrieve the embeddings of the words corresponding to each element in triplet t (i.e. person, time and location) from $\mathbf{X}_c$ to obtain t's representation $\mathbf{H}_t \in \mathbb{R}^{3 \times d}$.

After that, $\mathbf{H}_c$ is independently passed through three parallel Conv2d layers, each with their respective learnable weights. Let $\mathbf{h}_c^i, i \in \{1, 2, 3\}$ represent the outputs from each Conv2d layer. On the other hand, $\mathbf{H}_t$ undergoes a single Conv2d operation to get $\mathbf{h}_t$. Then we concatenate them to obtain $\hat{\mathbf{h}_{cnn}}$ (Eq. 2-3). Note that all the above Conv2d Layers have the same output dimension.

$$\mathbf{h}_* = \text{Conv2d}(\mathbf{H}_*), \tag{2}$$

$$\hat{\mathbf{h}_{cnn}} = \mathbf{h}_c^1 \oplus \mathbf{h}_c^2 \oplus \mathbf{h}_c^3 \oplus \mathbf{h}_t, \tag{3}$$

Finally, we get the representation $\mathbf{h}_{cnn}$ from CNN through a linear layer (Eq. 4).

$$\mathbf{h}_{cnn} = \mathbf{w}_c^\top \hat{\mathbf{h}_{cnn}} + \mathbf{b}_c, \tag{4}$$

where $\mathbf{w}_c$ is the weight parameter.

Representation from BERT We put t at the beginning of c, forming a complete input sequence s. The resulting s is then fed into the BERT model for fine-tuning, obtaining a combination representation $\mathbf{h}_{bert}$ (Eq. 5).

$$\mathbf{h}_{bert} = \text{BERT}(s). \tag{5}$$

Joint Training of CNN and BERT To jointly train CNN and BERT, we first transform $\mathbf{h}_{cnn}$ and $\mathbf{h}_{bert}$ through two independent learnable linear layers with the same output dimension k_2, resulting in $\mathbf{h}_{cnn}^{'}$ and $\mathbf{h}_{bert}^{'}$, respectively. Then, $\mathbf{h}_{cnn}^{'}$ and $\mathbf{h}_{bert}^{'}$ are concatenated into $\mathbf{h}_{ce}$ for classification (Eq. 6-9).

$$\mathbf{h}_{cnn}^{'} = \mathbf{w}_{cnn}^\top \mathbf{h}_{cnn} + \mathbf{b}_{cnn}, \tag{6}$$

$$\mathbf{h}_{bert}^{'} = \mathbf{w}_{bert}^\top \mathbf{h}_{bert} + \mathbf{b}_{bert}, \tag{7}$$

$$\mathbf{h}_{ce} = \mathbf{h}_{cnn}^{'} \oplus \mathbf{h}_{bert}^{'}, \tag{8}$$

$$y_i^{'} = \text{Softmax}(\mathbf{h}_{ce}), \tag{9}$$

where $\mathbf{w}_*$ is the weight parameter.

Additionally, we also preserve a combination representation for calculating the contrastive learning term (see the next section). Specifically, $\mathbf{h}_{cnn}$ and $\mathbf{h}_{bert}$ are individually passed through another two learnable linear layer with the same output dimension k_3, resulting in $\mathbf{h}_{cnn}^{\diamond}$ and $\mathbf{h}_{bert}^{\diamond}$, respectively. These representations are then concatenated and passed through an attention layer to obtain a combined feature vector $\mathbf{h}_{scl}$ (Eq. 10).

$$\mathbf{h}_{scl} = \text{attn}(\mathbf{h}^{\diamond}_{cnn} \oplus \mathbf{h}^{\diamond}_{bert}), \tag{10}$$

where attn is the attention layer.

Supervised Contrastive Learning Loss

Traditional classification tasks usually train classification models only with cross-entropy loss (Khosla et al. 2020). However, cross-entropy loss can lead to instability and poor generalization when labeled data is limited (Zhang and Sabuncu 2018). Supervised contrastive learning is proposed to combine the idea of "learn to compare" from contrastive learning into the supervised setting, alleviating the shortcomings of cross-entropy loss overly relying on the label itself. (Khosla et al. 2020).

With the output feature vector $\mathbf{h}_{scl}$, we next compute the supervised contrastive learning loss term in Eq. 11, following Gunel et al. (2020).

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} \frac{-1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i = y_j} \log \frac{e_{ij}}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} e_{ik}},$$
$$e_{mn} = \exp(h^m_{scl} \cdot h^n_{scl} / \tau), \tag{11}$$

where N is the batch size, h_{scl} is the feature vector from the ensemble model, N_{y_i} is the total number of examples in the batch that have the same label as y_i, and τ is an adjustable temperature parameter. We combine cross-entropy loss with $\mathcal{L}_{SCL}$ to calculate the loss $\mathcal{L}_S$ on labeled data with a scalar weighting hyperparameter λ (Eq. 12-13).

$$\mathcal{L}_S = (1 - \lambda)\mathcal{L}_{CE} + \lambda \mathcal{L}_{SCL}, \tag{12}$$

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i}^{N} (y_i \log(y_i^{'}) + (1 - y_i) \log(1 - y_i^{'})). \tag{13}$$

Semi-supervised Learning Loss

Semi-supervised learning is a learning paradigm that utilizes unlabeled data to enhance the performance of the model, particularly in scenarios where there is limited labeled data but a large amount of unlabeled data (Van Engelen and Hoos 2020). In our scenario, labeling triplets is time-consuming and difficult to ensure a sufficient number of long-tail trajectory contexts, while the triplet extraction pipeline can quickly generate large amounts of unlabeled data. Therefore, we introduce semi-supervised learning during training to enhance the generalization ability of the model through unlabeled data.

Pseudo-label (Lee et al. 2013) is a simple but efficient method performing semi-supervised learning for deep neural networks. The principle behind Pseudo-label is to take the model's prediction on unlabeled data as pseudo-labels, and compute the corresponding loss of unlabeled data. This loss is then combined with the loss computed on labeled data. We adopt the idea of Pseudo-label to shift our model from supervised to semi-supervised setting by computing $\mathcal{L}_U$. Here we use cross-entropy loss instead of the combination of $\mathcal{L}_{SCL}$ for unlabeled data, since the predicted label may not ensure that both positive and negative labels appear in each batch, which is the prerequisite for calculating $\mathcal{L}_{SCL}$.

$$\mathcal{L} = \mathcal{L}_S + \alpha(b, t)\mathcal{L}_U, \tag{14}$$

$$\mathcal{L}_U = \mathcal{L}_{CE}, \tag{15}$$

$$\alpha(b, t) = \begin{cases} 0 & b \leq c_1 B \\ \frac{b}{B} \cdot \frac{\gamma}{t+1} & c_1 B < b \leq c_2 B \\ 1 & b > c_2 B \end{cases}. \tag{16}$$

where b and t are ordinal numbers meaning the b-th batch in the t-th epoch, B is the number of total batches in each epoch, c_1, c_2, and γ are hyperparameters. $\alpha(b, t)$ is used to balance $\mathcal{L}_U$ and $\mathcal{L}_S$.

Experiments

Train/Test Split

We divide the "Representative" dataset into training and testing in a 7:3 ratio. Given the relatively independent nature of the annotated triplets obtained through sampling, a random split is less likely to result in data leakage. Additionally, we randomly sampled 20% from the training set as a validation set for hyperparameter tuning. Meanwhile, we preserve the "Regular" dataset as an independent test set.

Implementation Details

In this paper, we use a BERT-base[5] model to generate word embeddings, while any suitable model can be employed as well. In COSMOS, we set $d = 768$, $k_1 = 100$, $k_2 = 2$, and $k_3 = 32$. We set $\lambda = 0.2$, $\tau = 0.1$, $c_1 = 0.1$, $c_2 = 0.9$, and $\gamma = 0.8$ in loss terms. The hyperparameters including λ, τ, and γ are adjusted with grid search (Bergstra et al. 2011).

To train COSMOS, we set the learning rate to $5e^{-5}$ and use the Adam optimizer (Kingma 2014). Moreover, we adopt an early stop strategy to avoid overfitting. We conduct our experiments with two RTX 3090.

Evaluation Metrics

To quantitatively evaluate our model, we use the following performance metrics.

Metrics for Prediction Performance We use Accuracy (Acc), Precision (P), Recall (R), and F1 score to evaluate the model's ability to extract trajectory triplets by test set from "Representative". The F1 score holds significant importance in our scenario as it represents the balance of Precision and Recall, which has a great impact on downstream analysis. Given the mixed origins of our dataset samples (manually labeled and GPT-labeled), we denote the respective sub-dataset with subscripts for distinction. "Representative" refers to all the samples from both sources, "Representative$_m$" denotes manually labeled samples, and "Representative$_g$" represents GPT-labeled samples. Given that GPT is only used for labeling positive samples, we only calculate the Recall for GPT-labeled samples.

[5]https://github.com/google-research/bert

Metrics for Coverage Performance We use Recall (R) to assess the coverage performance of the model on all the trajectories within each biography page by "Regular". Additionally, we compute the average Recall and its standard deviation across different pages.

Baseline Methods

Here we introduce seven baselines. The first four models, LR (TF-IDF), CNN, Bi-LSTM, and CeleTrip, as mentioned in Related Work, are used in the previous work to extract spatio-temporal knowledge about people from text corpus and related extraction tasks. The next three, BERT, RoBERTa, and GPT-3.5, are language models commonly used for general language tasks. Recently, GPT-3.5 has gained significant attention as a general-purpose language model and has been evaluated in various information extraction tasks (Han et al. 2023; Gao et al. 2023). However, these studies have indicated that there is still a gap between the performance of GPT and the supervised SOTA in information extraction tasks such as Event Detection.

Since we use both the triplet and its context in COSMOS, as a fair comparison, we combine the triplet and its context as an entire input to each baseline.

- **LR (TFIDF) (Dickinson et al. 2015):** The researchers use TFIDF vectors for text representation on their classification task. We use a similar approach to extract representations and employ Logistic Regression as the classifier in our experiments.
- **CNN (Nguyen et al. 2017):** During their research into identifying crisis events from personal tweets, the model captures the local semantic features of text. The resulting features are then combined together and utilized as the input for the classifier. In our experimental setup, we employ the softmax classifier for all deep learning models.
- **Bi-LSTM (Yen, Huang, and Chen 2019):** When detecting real-life events from users' tweets, the researchers use Bi-LSTM to capture the sequential semantics within tweets content. In our task, we employ Bi-LSTM to model the triplets and their contexts.
- **CeleTrip (Peng et al. 2024):** The researchers propose CeleTrip for detecting celebrity itineraries from news articles. CeleTrip models the location context as a word graph and utilizes Oriented Pooling to encode information of the target celebrity and location. Similarly, we model triplet's context as a graph and fuse triplet features through Oriented Pooling in our experiments.
- **BERT (Devlin et al. 2018):** We consider BERT as part of our baselines. BERT is a widely used pre-trained language model and has demonstrated strong performance across various language tasks. When fine-tuning BERT for classification, we follow the same hyperparameter suggested by Devlin et al. (2018).
- **RoBERTa (Liu et al. 2019):** As an enhancement of BERT, RoBERTa is trained with longer data and more data while removing the Next Sentence Prediction (NSP) objective. It has been shown to outperform BERT on many downstream tasks (Liu et al. 2019). In the experiment, we use the same settings as BERT to fine-tune it.
- **GPT-3.5[6]:** We introduce GPT-3.5 (gpt-3.5-turbo-0613) as another baseline for our task. See the Appendix for the prompt we provide for GPT.

Additionally, we conduct partial experiments on GPT-4[7] (gpt-4-0613) due to the cost of API call, and the results and implementation details are reported in the Appendix.

Experimental Results

We evaluate the performance of models from two aspects. The "Representative" dataset is used to assess the prediction performance of models on sampled trajectory triplets, while the "Regular" dataset evaluates the model's coverage performance on life trajectories within complete biography pages.

Prediction Performance The results of each model on datasets "Representative" and "Regular" are shown in Table 1. Among these models, COSMOS achieves the best overall performance (F1=85.95%), followed by RoBERTa.

As the model with the best overall performance, COSMOS integrates CNN and BERT, approaching CNN in Precision and approaching BERT in Recall, indicating its effective fusion of the advantages of both models. The CNN model, which focuses on local semantics, achieves the best Precision (84.91%), while the graph-based model CeleTrip, which captures long-distance semantics, has a lower performance (81.77%). The possible reason may be that in our scenario, the text is relatively short, and the semantics are more concentrated, making sequential models more suitable for text modeling than graph-based models.

Additionally, we notice that although the GPT-3.5 model achieves the highest Recall (95.12%), it exhibits a severe imbalance between Precision (56.53%) and Recall, introducing a significant amount of noise while retrieving life trajectories. Given its tendency to misinterpret temporal and spatial information in the text as trajectories, it is not yet suitable for direct application in extracting life trajectories. Meanwhile, BERT, which focuses on contextual understanding, achieves the highest Recall (88.80%) among models other than GPT-3.5. RoBERTa has a somehow unexpected lower Recall compared to BERT (decreased by 0.76%), while overall outperforming BERT in our task (increased by 1.15% in F1). Different from BERT, RoBERTa's design removes the NSP Loss during the training process, which might have an impact on downstream tasks that require language inference (Devlin et al. 2018). The Bi-LSTM model, which also captures bidirectional semantic information like BERT and RoBERTa, achieves relatively higher Recall than other models. This indicates that the capturing of continuous semantics contributes to retrieving more trajectories in our scenario.

The performance of these models on "Representative$_m$" is similar to those on "Representative". However, we observe an intriguing phenomenon in our experimental results, where all models except LR (TFIDF) exhibit a superior increase on Recall when tested on the samples labeled by GPT-3.5 ("Representative$_g$"). To explore the possible reason for

[6]https://platform.openai.com/docs/models/gpt-3-5-turbo

[7]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

	Representative				Representative$_m$				Representative$_g$	Regular	
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)	R (%)	R (%)	Avg-R (std)
GPT-3.5	63.99	56.53	95.12*	70.91	55.00	41.48	91.39*	57.06	100.00*	92.33*	0.9126 ± 0.0716
LR (TFIDF)	74.47	75.45	66.24	70.55	75.67	62.62	63.64	63.13	69.64	44.52	0.4262 ± 0.1751
CeleTrip	82.55	81.77	80.05	80.90	81.31	70.26	74.33	72.24	87.54	60.94	0.5614 ± 0.2351
Bi-LSTM	84.38	81.38	85.77	83.52	81.94	69.66	79.37	74.20	94.16	75.18	0.7549 ± 0.2031
CNN	84.42	**84.91**	80.55	82.67	82.62	**74.08**	72.10	73.08	91.63	63.50	0.6344 ± 0.2111
BERT	84.65	80.10	**88.80**	84.23	82.08	68.39	**84.12**	75.44	94.94	81.02	0.8304 ± 0.1398
RoBERTa	86.09	82.88	88.04	85.38	83.68	71.94	82.19	76.73	**95.71**	77.00	0.7389 ± 0.1583
COSMOS	**86.79**	84.41	87.54	**85.95**	**84.61**	**74.08**	81.45	**77.59**	95.52	**82.11**	0.8169 ± 0.0906

Table 1: Performance comparison on the test set. Due to the extreme imbalance between Precision and Recall of GPT-3.5, we specifically highlight the Recall for it with an asterisk (*). Apart from that, the best results are indicated by bold text, while the second-best ones are highlighted with underlines.

that, we compare these two datasets ("Representative$_m$" and "Representative$_g$") by computing the distribution of verbs from the positive triplets' contexts. The top 5 most frequent words ("born", "be", "die", "move" and "marry") appear in the exact same order in both datasets, and accounts for 40.20% and 44.55%, respectively. Then, we present the distribution of verbs from 6th to 10th in Table 2. It seems that manually annotated data involves more personal life and behavior, while GPT-3.5 annotated data focuses more on work, responsibilities and performance. This raises the question of why models using pre-trained word embeddings display consistently better learning outcomes on the GPT-3.5 annotated data, despite there being no significant difficulty difference in verbs between the two test sets. Further investigation into the underlying factors influencing this phenomenon and its generalizability across different tasks and datasets presents an interesting direction for future research.

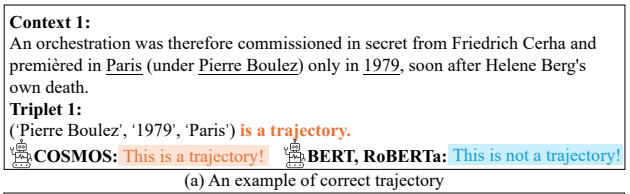
Representative$_m$	Representative$_g$
graduate (2.5%)	perform (2.3%)
become (2.2%)	make (2.3%)
live (2.0%)	hold (1.5%)
take (1.6%)	work (1.5%)
leave (1.6%)	serve (1.5%)

Table 2: Top 6-10 high-frequency verbs in triplets' contexts.

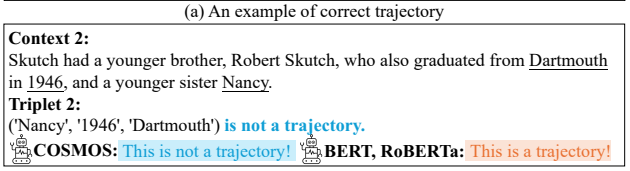
Coverage Performance We evaluate the models' coverage of life trajectories on individual Wikipedia biography pages using the "Regular" dataset. Most models have lower Recall on "Regular" dataset compared to the Recall on manually labeled samples in "Representative" dataset.

Apart from GPT-3.5, only COSMOS shows a slight improvement (increased by 0.66%) in this metric. This indicates that the combination of contrastive learning and semi-supervised learning enhances the model's generalization ability. Additionally, excluding GPT-3.5, which sacrifices Precision for Recall, COSMOS achieves the highest Recall (82.11%) and exhibits the least standard deviation across different pages. This suggests that COSMOS is practical in real-world applications.

Examples of Trajectories Extracted Alongside the quantitative analysis, we also examine the trajectories extracted by different models. To illustrate COSMOS's capacity for generalization across diverse trajectory descriptions and its precision in extracting information, we provide two representative examples below. Figure 4 (a) is the trajectory of Pierre Boulez conducting premiere in Paris, described as *"an orchestration ... (under Pierre Boulez)"*. COSMOS gives the correct classification, while BERT and RoBERTa miss this relatively implicit description of trajectory. Meanwhile, Figure 4 (b) is an example of an incorrect trajectory, and the given triplet is *("Nancy", "1946", "Dartmouth")*. COSMOS identifies the triplet as not being a trajectory of *"Nancy"* while the other two models seem to be misled by other people's trajectory.



(a) An example of correct trajectory



(b) An example of incorrect trajectory

Figure 4: Examples of extracted trajectories. Underlines represent the corresponding element in the given triplet.

Error Analysis We perform an error analysis on COSMOS to analyze its limitations. Overall, there are two types of error: false positives (triplets that are mistaken for trajectory, related to Precision) and false negatives (trajectory triplets that are missed, related to Recall). We sample 20% from each type, resulting in 34 false positives and 35 false negatives.

The main error reasons for false positives are (1) the lack of background knowledge (59%) and (2) mismatched time (18%). We demonstrate reason (1) by the following example from one biography page: *"Viceroy of India, Lord Curzon, partitioned Bengal."* The verb *"partition"* in this context means the allocation of resources rather than implying that *"Lord Cornwallis"* himself was present in *"Bengal"*; this may require a good grasp of geopolitical knowledge to help the model understand its underlying meaning. For reason (2), time expressions included in parentheses such as *"1887"* from *"the Artist's Wife (I Havedøren, 1887)"* can mislead the model in finding the actual date of trajectory.

False negatives are mainly caused by: (1) vague time explanation (29%); and (2) diversity in writing style (26%). Time expressions such as *"monthly"* and *"the next few years"* sometimes distract the model from finding the trajectories. On the other hand, complex sentences written by diverse editors (Ren and Yan 2017) such as *"Exactly the same injury, 44 years later, in **August 1986**, in **Afghanistan**, his grandson, a military intelligence sergeant, **Iliyas Daudi**, who was blown up by an Italian anti-personnel mine."* are challenging. The writing style with numerous comma-separated phrases makes it difficult for the model to connect essential information and classify it as a positive trajectory.

Ablation Study

We remove certain loss terms from COSMOS to validate their effectiveness, and the results are shown in Table 3. COSMOS $_{\text{w/o ssl}}$ removes the semi-supervised learning loss. Similarly, COSMOS $_{\text{w/o scl}}$ removes the supervised contrastive learning loss. COSMOS $_{\text{w/o ssl\&scl}}$ removes both the semi-supervised learning and supervised contrastive learning loss terms.

As shown in Table 3, COSMOS demonstrates the best overall performance as expected, while COSMOS $_{\text{w/o ssl\&scl}}$ performs the worst, indicating the effectiveness of the corresponding design. Compared to COSMOS $_{\text{w/o ssl\&scl}}$, the inclusion of contrastive learning (COSMOS $_{\text{w/o ssl}}$) or semi-supervised learning (COSMOS $_{\text{w/o scl}}$) alone leads to an increase in Precision (increased by 2.64% and 4.47%, respectively) and a decrease in Recall (decreased by 2.19% and 2.61%, respectively), but improving the overall performance of the model. However, when both are incorporated into the model, Precision and Recall achieve a better balance, and result in a significant improvement in coverage performance (increased by 13.14% in Recall on "Regular"). It is possible that the abilities of contrastive learning and semi-supervised learning are complementary. Contrastive learning enhances the discriminative ability of the model by learning the similarity between samples (Khosla et al. 2020), while semi-supervised learning provides additional information by utilizing unlabeled data (Van Engelen and Hoos 2020). When they are combined, contrastive learning facilitates the selection of more discriminative features, while semi-supervised learning can expand the training set and increase data diversity, which may contribute to enhancing the model's performance.

Analysis of a Sample Set

Trajectories of antiquarians, scientists, artists, and historians have been a focus on cultural history studies (Schich et al. 2014; Fu 2014; Kaiser et al. 2018; Long 2018). In this section, we extract and analyze the trajectories of historians, to further validate our method and provide a taste of our dataset.

Extracting and Processing Trajectories

We use the set of historians identified by Laouenan et al. (2022) and retrieve their biography pages from Wikipedia. After extracting candidate triplets and classifying them by COSMOS, we clean and augment the extracted trajectories by utilizing Hugging Face's co-reference resolution tool[8] for disambiguating the names. Additionally, we standardize time information and geocode location information using Nominatim[9].

Data Quality, Density and Variety

After data processing, we obtain 20,786 trajectory triplets of 8,272 historians. We manually check 225 of the trajectory triplets and find that 80% of them (180) are accurate, which is comparable with the model performance on the test set. On average, each person has 2.51 triplets. We further examine the types of trajectory triplets in this sample set. Similar to Lucchini, Tonelli, and Lepri (2019), we use the verbs near the trajectory triplets to roughly represent the types according to the classification in FrameNet[10] and Figure 5 shows the distributions of the top 15 verbs. Other than births and deaths, activities such as education, work, moving and travel take up 81.25%, indicating that various interim points can add many new dimensions to datasets with only births and deaths. Figure 6 showcases the trajectories of three historians, *H. Bruce Franklin* (5 triplets, in the 20th century), *John Henry Brown* (6 triplets, in the 19th century), and *Karl Theodor Keim* (2 triplets, in the 19th century). We can observe different movement patterns – *Keim*'s activities seem to have a limited geographic span, while *Franklin* is an active traveler across continents, which may have benefited from the convenience of long-distance travel in the second half of the 20th century. *Brown* had six moves within the U.S. and Mexico between 1845 and 1885, spending most of his life in Texas. The movements can surely be further quantified and aggregated over more historians to seek meaningful insights.

Spatio-Temporal Interaction Network We construct the interaction network based on the temporal and spatial intersections of historians' trajectories, which can not be easily created from other datasets. We restrict the location must contain words related to schools and institutes such as University, College, etc. to ensure that the interactions are most likely to have happened, instead of merely capturing coarse spatio-temporal co-occurrences. Overall, the network contains 899 nodes (historians) and 791 edges (interactions),

[8] https://github.com/huggingface/neuralcoref

[9] https://nominatim.org/

[10] https://framenet.icsi.berkeley.edu/

	Representative				Representative$_m$				Representative$_g$	Regular	
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)	R (%)	R (%)	Avg-R (std)
COSMOS $_{w/o\ ssl\&scl}$	85.23	83.00	85.52	84.24	82.66	71.62	77.89	74.62	**95.52**	68.97	0.6955 ± 0.1791
COSMOS $_{w/o\ ssl}$	85.85	85.64	83.33	84.47	83.83	75.33	75.22	75.27	93.96	69.34	0.6636 ± 0.2479
COSMOS $_{w/o\ scl}$	86.63	**87.47**	82.91	85.13	**84.80**	**78.07**	74.48	76.23	93.96	71.89	0.6777 ± 0.2109
COSMOS	**86.79**	84.41	**87.54**	**85.95**	84.61	74.08	**81.45**	**77.59**	**95.52**	**82.11**	0.8169 ± 0.0906

Table 3: Results of the ablation study. Bold text indicates the best results, while underlined text represents the second-best ones.

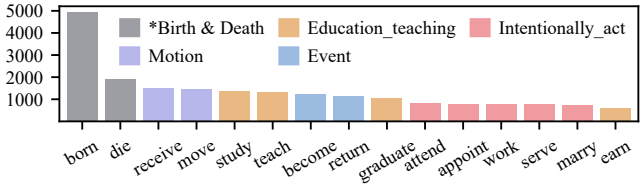


Figure 5: The distribution of the top 15 frequent verbs associated with the trajectories of historians. The horizontal axis represents verbs and the vertical axis represents their corresponding quantities. The * legend indicates the custom category independent of FrameNet.

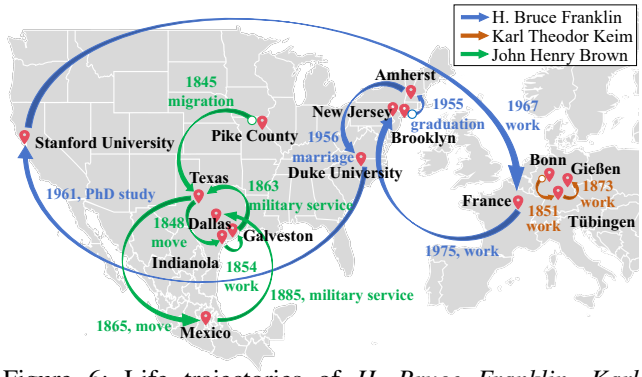


Figure 6: Life trajectories of *H. Bruce Franklin*, *Karl Theodor Keim* and *John Henry Brown*. The arrows of each color represent the life trajectory of the corresponding individual. The start point of each trajectory is marked with a circle. The year and purpose of the move are labeled on the arrows.

spanning from 1811 to 2019. In Figure 7, we see graph snapshots taken every 10 years from 1910 to 2020. Node sizes are calculated based on PageRank and the colors indicate the nationalities of the historians. It seems that from 1940 to 1970, there is a rapid increase in nodes and edges, compared with other periods of time. As a comparison, from 2000 to 2020, the increase appears to have slowed down. These are further supported by the distribution of historians' birth years, which are concentrated around 1920-1940. We suspect that similar dynamics may also be reflected in the publication/citation records of history papers. Subplots (a) and (b) demonstrate two connected components from a microscopic point of view and from the colors we can see that these interactions mostly happen between historians from the same coun-

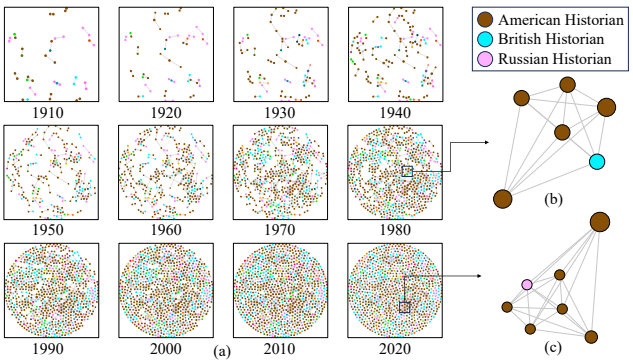


Figure 7: Dynamic interaction network comprising 899 historians. (a) Snapshots of the network every 10 years from 1910 to 2020. Nodes represent historians, the sizes of nodes are the PageRank values, and their nationalities are indicated by colors. The visualization is created using the Fruchterman Reingold layout. (b) and (c) zoom in on two connected components in the 1980 snapshot and 2020 snapshot respectively.

tries.

Beyond these analytics, the popular Spatio-Temporal Graph Convolutional Networks (Yu, Yin, and Zhu 2017; Hu et al. 2023; Lian et al. 2023) may help better model the graph and provide predictive insights.

As an anecdote, *William H. McNeill* from the University of Chicago, who died at the age of 98, had the highest PageRank score in this interaction graph.

Conclusion and Future Work

We propose a new task of extracting life trajectories from Wikipedia and introduce COSMOS, to effectively extract life trajectories from Wikipedia biography pages by combining the idea of contrastive learning and semi-supervised learning. To validate the method and showcase the potential of the extracted data, we extract the trajectories of historians, and perform an analysis based on the resulting trajectories. We hope the open-sourced code, the million-level extracted trajectories, and the *WikiLifeTrajectory* ground truth dataset, can support the trajectory extraction research and the analytical studies based on these trajectories. As the largest of its kind, our dataset can be the basis for data-driven grand narratives and explorations of human mobility and interaction mechanisms. Beyond births and deaths, this compre-

hensive compilation encompasses various life milestones, offering insights into aspects such as education, work, and marriage. All our data shared from this work will be made FAIR (FORCE11 2020).

We have to note that since we choose to extract trajectories from the English Wikipedia, there can be a bias that the extracted people are more likely to be from the English world (Roy, Bhatia, and Jain 2021). This should be considered when any research tries to draw conclusions from our dataset. To mitigate this, a possible future step is to extend our framework to versions of Wikipedia in other languages and further explore different designs of extraction algorithms. As another future improvement, we may include identifying triplet types such as "graduate study", "attending a conference", and "delivering a speech" in our task, to extract the purposes of trajectories.

Ethics Statement

Our study uses the English Wikipedia, and we ensure that our data collection process does not violate any privacy or confidentiality concerns. A potential ethical concern is the misuse to extract the life trajectories of individual users (non-famous people). However, since our framework relies on a detailed description of one's life, the risk would arise from the leakage of personal information in such a description, rather than the framework itself. Therefore, we believe that there are no essential ethical questions raised by our study.

Acknowledgments

The authors thank the editors and anonymous reviewers for their invaluable assistance in improving the quality of the paper. Additionally, we would like to extend our gratitude to Yixi Zhou, Xiaoxia Zhang, and Hairui Yin, for their assistance during the data annotation process.

References

Babaiha, N. S.; Rao, S. G.; Klein, J.; Schultz, B.; Jacobs, M.; and Hofmann-Apitius, M. 2024. Rationalism in the face of GPT hypes: Benchmarking the output of large language models against human expert-curated biomedical knowledge graphs. *Artificial Intelligence in the Life Sciences*, 5: 100095.

Becker, G. S. 1974. A Theory of Social Interactions. *Journal of Political Economy*, 82: 1063 – 1093.

Bergstra, J.; Bardenet, R.; Bengio, Y.; and Kégl, B. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24.

Boianovsky, M. 2018. 2017 HES Presidential Address: Economists and their travels, or the time when JFK sent Douglass North on a mission to Brazil. *Journal of the History of Economic Thought*, 40(2): 149–177.

Deville, P.; Wang, D.; Sinatra, R.; Song, C.; Blondel, V. D.; and Barabási, A.-L. 2014. Career on the move: Geography, stratification and scientific impact. *Sci. Rep*, 4(1): 4770.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dickinson, T.; Fernandez, M.; Thomas, L. A.; Mulholland, P.; Briggs, P.; and Alani, H. 2015. Identifying prominent life events on twitter. In *Proceedings of the 8th International Conference on Knowledge Capture*, 1–8.

Doherty, B. J. 2009. POTUS on the Road: International and Domestic Presidential Travel, 1977-2005. *Presidential Studies Quarterly*, 39(2): 322–346.

Elder, G. H.; Johnson, M. K.; and Crosnoe, R. 2003. *The emergence and development of life course theory*. Springer.

Elder Jr, G. H. 1994. Time, human agency, and social change: Perspectives on the life course. *Social psychology quarterly*, 4–15.

Foppiano, L.; Lambard, G.; Amagasa, T.; and Ishii, M. 2024. Mining experimental data from Materials Science literature with Large Language Models. *arXiv preprint arXiv:2401.11052*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-04-02.

Fu, M. 2014. A cultural analysis of China's scientific brain drain: The case of Chinese immigrant scientists in Canadian academia. *Journal of International Migration and Integration*, 15: 197–215.

Gao, J.; Zhao, H.; Yu, C.; and Xu, R. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Gebhard, L.; and Hamborg, F. 2020. The POLUSA dataset: 0.9 M political news articles balanced by time and outlet popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 467–468.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Goldsmith, B. E.; Horiuchi, Y.; and Matush, K. 2021. Does public diplomacy sway foreign public opinion? Identifying the effect of high-level visits. *American Political Science Review*, 115(4): 1342–1357.

Gunel, B.; Du, J.; Conneau, A.; and Stoyanov, V. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Han, R.; Peng, T.; Yang, C.; Wang, B.; Liu, L.; and Wan, X. 2023. Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors. *arXiv preprint arXiv:2305.14450*.

Hautala, J.; and Jauhiainen, J. S. 2019. Creativity-related mobilities of peripheral artists and scientists. *GeoJournal*, 84(2): pp. 381–394.

Hu, W.; Li, W.; Zhou, X.; Kawai, A.; Fueda, K.; Qian, Q.; and Wang, J. 2023. Spatio-Temporal Graph Convolutional Networks via View Fusion for Trajectory Data Analytics. *IEEE Trans. Intell. Transp. Syst.*, 24(4): 4608–4620.

Kaiser, M.; Lejtovicz, K.; Schlögl, M.; and Rumpolt, P. A. 2018. Artist migration through the biographer's lens: A case

study based on biographical data retrieved from the Austrian Biographical Dictionary. *Journal of Historical Network Research*, 2: 76–108.

Kernell, S. 2006. *Going public: New strategies of presidential leadership*. Cq Press.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NIPS*, 33: 18661–18673.

Kingma, D. 2014. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*.

Kleinepier, T.; de Valk, H.; and van Gaalen, R. 2015. Life Paths of Migrants: A Sequence Analysis of Polish Migrants' Family Life Trajectories. *European Journal of Population*, 31: 155 – 179.

Laouenan, M.; Bhargava, P.; Eyméoud, J.-B.; Gergaud, O.; Plique, G.; and Wasmer, E. 2022. A cross-verified database of notable people, 3500BC-2018AD. *Sci. Data*, 9(1): 290.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML*, volume 3, 896. Atlanta.

Li, P.; Yao, J.; Wang, L.; and Lin, X. 2017. SPOT: Selecting occuPations from trajectories. In *SIGIR*, 813–816.

Lian, J.; Ren, W.; Li, L.; Zhou, Y.; and Zhou, B. 2023. Ptp-stgcn: pedestrian trajectory prediction based on a spatio-temporal graph convolutional neural network. *Applied Intelligence*, 53(3): 2862–2878.

Liu, Y.; Hua, W.; and Zhou, X. 2019. Extracting Temporal Patterns from Large-Scale Text Corpus. In *Australasian Database Conference*.

Liu, Y.; Hua, W.; and Zhou, X. 2020. Temporal knowledge extraction from large-scale text corpus. *World Wide Web*, 24: 135 – 156.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Long, Q. 2018. A Comparison between the Methodology of Recording History of Ancient Historians. *ICCLAH*.

Lucchini, L.; Tonelli, S.; and Lepri, B. 2019. Following the footsteps of giants: modeling the mobility of historically notable individuals using Wikipedia. *EPJ Data Science*, 8(1): 36.

Möller, C.; Lehmann, J.; and Usbeck, R. 2022. Survey on english entity linking on wikidata: Datasets and approaches. *Semantic Web*, 13(6): 925–966.

Morrissey, R. M. 2015. Archives of Connection. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 48: 67 – 79.

Nguyen, D.; Al Mannai, K. A.; Joty, S.; Sajjad, H.; Imran, M.; and Mitra, P. 2017. Robust classification of crisis-related data on social networks using convolutional neural networks. In *ICWSM*, volume 11, 632–635.

Peng, K.; Zhang, Y.; Ling, S.; Ke, Z.; and Zhang, H. 2024. Where Did the President Visit Last Week? Detecting Celebrity Trips from News Articles. In *ICWSM*, volume 18, 1193–1206.

Piskorski, J.; Zavarella, V.; Atkinson, M.; Verile, M.; et al. 2020. Timelines: Entity-centric Event Extraction from Online News. In *Text2Story@ ECIR*, 105–114.

Ren, R.; and Yan, B. 2017. Crowd diversity and performance in wikipedia: The mediating effects of task conflict and communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6342–6351.

Roy, D.; Bhatia, S. K.; and Jain, P. 2021. Information asymmetry in Wikipedia across different languages: A statistical analysis. *JASIST*, 73: 347 – 361.

Schatzki, T. R. 2019. *Social change in a material world*. Routledge.

Schatzki, T. R. 2022. The trajectories of a life. In *Doing Transitions in the Life Course: Processes and Practices*, 19–34. Springer International Publishing Cham.

Schich, M.; Song, C.; Ahn, Y.-Y.; Mirsky, A.; Martino, M.; Barabási, A.-L.; and Helbing, D. 2014. A network framework of cultural history. *science*, 345(6196): 558–562.

Thapa, S.; Naseem, U.; and Nasim, M. 2023. From humans to machines: can ChatGPT-like LLMs effectively replace human annotators in NLP tasks. In *ICWSM*.

Van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2): 373–440.

Vempala, A.; and Blanco, E. 2020. Extracting biographical spatial timelines: corpus and experiments. *IEEE Trans Audio Speech Lang Process*, 28: 1395–1403.

Verginer, L.; and Riccaboni, M. 2020. Cities and countries in the global scientist mobility network. *Applied Network Science*, 5: 1–16.

Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; and Wang, G. 2023. GPT-NER: Named Entity Recognition via Large Language Models. *ArXiv*, abs/2304.10428.

Yen, A.-Z.; Chang, C.-C.; Huang, H.-H.; and Chen, H.-H. 2021. Personal knowledge base construction from multimodal data. In *SIGIR*, 496–500.

Yen, A.-Z.; Huang, H.-H.; and Chen, H.-H. 2019. Personal knowledge base construction from text-based lifelogs. In *SIGIR*, 185–194.

Yu, B.; Yin, H.; and Zhu, Z. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *NIPS*, 31.

Zucker, L. G.; and Darby, M. R. 2009. Star scientists, innovation and regional and national immigration. In *Entrepreneurship and Openness*. Edward Elgar Publishing.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Introduction, Ethics Statement, Conclusion and Future Work.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see the Abstract and Introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see the Introduction, Experimental Results, Conclusion and Future Work.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Conclusion and Future Work.
   (e) Did you describe the limitations of your work? Yes, see the Conclusion and Future Work.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes, see the Ethics Statement.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Ethics Statement. The data we use is from Wikipedia, a publicly available website.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see the Experimental Results.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, see the Dataset. We use stratified sampling according to occupation to avoid potential bias.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, see the Analysis of a Sample Set, Conclusion and Future Work.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see the Introduction, Analysis of a Sample Set, Conclusion and Future Work.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, see the open-sourced repository https://anonymous.4open.science/r/wiki_life_trajectory/.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, see the Experiments.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because our experiment is conducted using only one fixed random seed of 42 and controls all random numbers to ensure that our experimental results can be reproduced.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Implementation Details section in Experiments.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, see the Experimental Results section in Experiments.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, see the Experimental Results section in Experiments.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, see the Method.
   (b) Did you mention the license of the assets? No, because all the tools, algorithms and data we use are publicly available.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, see the URLs in footnotes.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see the Ethics Statement. The data we use is from Wikipedia, a publicly available website.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Ethics Statement. The data we use is from Wikipedia, a publicly available website.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, see the Conclusion and Future Work.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al.

(2021))? Yes, we follow the instructions and create a Datasheet.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Prompt for GPT-3.5

Prompt for Annotating Triplets Figure 8(a) and 8(b) show the given prompts when we use GPT-3.5 (gpt-3.5-turbo-0613) for data annotation. For each input sentence, we first extract triplets through the prompt of Figure 8(a), and then use the prompt of Figure 8(b) to perform self-verification. During the annotation step, we set the temperature (the parameter used to adjust the diversity of the model's output) to 0.8 and three trials are performed for each input. In the self-verification step, the temperature is set to 0 since we need GPT-3.5 to give stable final results.

Prompt for the Baseline Method Figure 8(c) shows the prompt provided when we use GPT-3.5 (gpt-3.5-turbo-0613) as the baseline. The temperature here is set to 0 and one trial is performed for each input.

Performance of GPT-4

To evaluate the performance of GPT-4 on the life trajectory extraction task, we test it on a quarter of the test set (643 instances random sampled from the "Representative" test set). During the experiments, we set the temperature to 0 and use the same prompt given to GPT-3.5 (see Figure 8(c)), with one trial for each input, which costs about $5 in total (around 250,000 tokens).

The experimental results are reported in Table 4. It is observed that GPT-4 outperforms GPT-3.5 in overall performance (increased by 8.33% in F1), while it is still weaker than most supervised methods. As general-purpose language models, ChatGPT-like LLMs might not yet be ready to be directly applied in large-scale extraction tasks instead of specialized models (Foppiano et al. 2024).

	Representative				**Representative$_m$**				**Representative$_g$**
	Acc (%)	P (%)	R (%)	F1 (%)	Acc (%)	P (%)	R (%)	F1 (%)	R (%)
GPT-4	79.63	75.53	83.33	79.24	76.56	61.61	76.92	68.42	91.60

Table 4: Results of GPT-4 baseline. From the "Representative" test set, 643 samples are randomly sampled (512 manually labeled, 131 GPT-3.5 labeled).

Additionally, we notice the Recall of GPT-4 is much lower than that of GPT-3.5 (decreased by 11.79%), which is also observed by other researchers when extracting biological information (Babaiha et al. 2024). Why GPT-based models produce such a phenomenon may be of interest to LLMs researchers.

Prompt for Extraction

Prompt:
Given a sentence, your task is to first detect person, time and location from the given sentence, and then determine whether the person was physically present in the specified location during the given time period. If all three elements exist and the relation between them is correct, return them in tuple-form, otherwise return None.
Input:
In 1975, Putin joined the KGB and trained at the 401st KGB School in Okhta, Leningrad.
Output: (Putin; 1975; Okhta, Leningrad)

Input:
In March 2015 Gentiloni visited Mexico and Cuba and met Cuban President Raúl Castro.
Output: (Gentiloni; March 2015; Mexico) (Gentiloni; March 2015; Cuba) (Raúl Castro; March 2015; Cuba)

Input:
Scranton fell into economic decline during the 1950s and Biden's father could not find steady work.
Output: None

Input:
Bob Hayes represented the USA in the 1964 Summer Olympics in Tokyo.

Expected Output: (Bob Hayes; 1964; USA)

(a)

Prompt for Self-Verification

Prompt:
Given a sentence, your task is to verify whether the given sentence specifies the person's physical presence during the given period, and whether this physical location is the given location.

The given sentence:
To support the album, Sheeran embarked on a world tour starting on 6 August 2014 at Osaka, Japan.
Does the given sentence specify Sheeran's physical presence in Osaka, Japan during 6 August 2014? Please answer with Yes or No.
Answer: Yes

The given sentence:
In October 2016, Macron criticized Hollande's goal of being a normal president, saying that France needed a more Jupiterian presidency.
Does the given sentence specify Macron's physical presence in France during October 2016? Please answer with Yes or No.
Answer: No

The given sentence:
Bob Hayes represented the USA in the 1964 Summer Olympics in Tokyo.
Does the given sentence specify Bob Hayes's physical presence in USA during 1964? Please answer with Yes or No.

Expected Output: No

(b)

Prompt for GPT3.5 Baseline

Prompt:
Given several sentences, your task is to verify whether these given sentences specify the person's physical presence during the given period, and whether this physical location is the given location.

The given sentences:
In 1989, Isaacs attended Musicians Institute in Hollywood, California, where he studied guitar. In late 1990, he left the Mad Hatter's Espresso Bar to host the open mike at Highland Grounds Café in Hollywood. Do the given sentences specify Isaacs is physically present in Hollywood, California during 1989? Please answer with Yes or No.
Answer: Yes

The given sentences:
In October 2016, Macron criticized Hollande's goal of being a normal president, saying that France needed a more Jupiterian presidency. Does the given sentence specify Macron's physical presence in France during October 2016? Please answer with Yes or No.
Answer: No

The given sentences:
Bob Hayes represented the USA in the 1964 Summer Olympics in Tokyo. Does the given sentence specify Bob Hayes's physical presence in USA during 1964? Please answer with Yes or No.

Expected Output: No

(c)

Figure 8: Prompt for GPT.