Fine-grained Classification of A Million Life Trajectories from Wikipedia

Anonymous Authors

Anonymous Institution

Abstract. Life trajectories of notable people convey essential messages for human dynamics research. These trajectories consist of (*person, time, location, activity type*) tuples recording when and where a person was born, went to school, started a job, or fought in a war. However, current studies only cover limited activity types such as births and deaths, lacking large-scale fine-grained trajectories. Using a tool that extracts (*person, time, location*) triples from Wikipedia, we formulate the problem of classifying these triples into 24 carefully-defined types using textual context as complementary information. The challenge is that triple entities are often scattered in noisy contexts. We use syntactic graphs to bring triple entities and relevant information closer, fusing them with text embeddings to classify life trajectory activities. Since Wikipedia text quality varies, we use LLMs to refine the text for more standardized syntactic graphs. Our framework achieves 84.5% accuracy, surpassing baselines. We construct the largest fine-grained life trajectory dataset with 3.8 million labeled activities for 589,193 individuals spanning 3 centuries. In the end, we showcase how these trajectories can support grand narratives of human dynamics across time and space. Code/data are publicly available.

Keywords: Life Trajectory · Human Dynamics · Wikipedia.

1 Introduction

Life trajectories [8] of notable figures capture their spatiotemporal activities, shedding light on human dynamics, technological shifts, idea diffusion, and innovation hubs [3,28]. A study published in *Science* [24] of over 300,000 individuals' birth/death locations reveals long-term cultural center evolution. Other research examines specific activities (e.g., scientists' careers, creative talent mobility, tech labor flows, artist or politician travels [11,10,20,25]) to understand socioeconomic and political developments. Recent LLM agents further enable simulation and interactive exploration of historical figures.

Prior data-driven effortss [24], however, have focused on easily accessible activity types—births/deaths, tech jobs [20], and scientific careers [11]—from sources like Freebase, LinkedIn, and academic papers [2]. Other important activities (e.g., political visits [2], career histories [31], marriages [33], education [23]) remain data-scarce and often require manual collection.

Although Wikipedia biographies offer rich sources for extracting *(person, time, location)* triples that represent trajectories, naively applying NER and combining candi-

date entities yields only around 30% precision [39]. Consequently, the COSMOS system was designed to extract such triples together with their contexts at high F1. This alleviates the data scale problem and provides a foundation for our work. However, obtaining fine-grained trajectory activity types remains an open challenge. We decide against formulating this as unsupervised learning, since clustering methods for these tasks [40,22] tend to produce uninterpretable results.

We formulate it into a multi-classification problem – given a triple of *(person, time, location)* and its context sentence from Wikipedia (i.e., the output of the COSMOS extraction tool [39]), decide the type of its trajectory activity. For example, as shown in Fig. 1, given the triple ('He', '1905', 'Adelaide') and its corresponding context sentence (a), our objective is to classify the type of this specific life trajectory activity (career). As a necessary pre-step, we draw upon the types of ACE [7] event extraction dataset and related work of life trajectory analysis [10,11,37], and propose a taxonomy including **24 types of life trajectory activities in 9 categories**.

This task is different from standard text classification [18] by classifying triples within their sentential context, requiring joint consideration of both. The triple must be emphasized and the semantic relations between its entities and relevant text aggregated. Without the triple, the sentence alone (e.g., Fig. 1) does not clearly indicate the activity type; only by situating the triple in the sentence can the relevant trajectory information be identified. This constraint resembles aspect-level sentiment classification (ALSC), which classifies sentiment for a given entity. ALSC methods often use syntactic graphs to capture associations between sentiment words and target entities, measuring graph distances. Inspired by this, we apply a syntactic graph to shorten the average pairwise distance among triple entities [5,38] (e.g., from 8.33 to 3 words), thereby reducing noise by focusing on words along the shortest paths. However, unlike ALSC approaches that merely concatenate textual representations with syntactic embeddings without fusion during training, we emphasize effective integration.

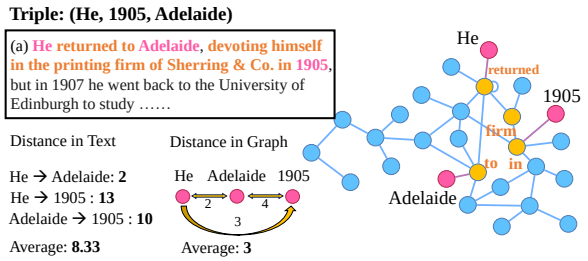

Fig. 1: Examples of life trajectories extracted from Wikipedia. Pink words are the extracted entities in triples indicating **person, time and location**. The graph is the syntactic graph constructed by SpaCy. The red nodes correspond to entities in triples. The orange ones are nodes and words on the paths.

Attention [27] aims to focus on the most relevant information in the entire sequence and its MASK [6] helps focus on the important words by masking words that do not matter. This inspires us to transform the syntactic structure of a sentence into a MASK vector that guides language models to associate textual pieces according to how they are structurally related and output text embedding fused with syntactic structure. As mentioned previously, the words in the shortest paths among the triple entities on the

syntactic graph are the ones we want to emphasize, forming a local **subgraph** of the overall syntactic **graph**. This subgraph's corresponding locations are 'unmasked' in MASK, guiding the text embedding to focus on the triple and its relevant information. Though these embeddings represent the words connected to the triple entities, other words in the sentence are not included. We therefore combine the regular textual embedding to obtain a more comprehensive representation of a life trajectory activity.

We note that collaboratively edited Wikipedia, lacks a consistent writing style, as observed in other studies [17]. This results in significant variation in syntactic structures, even for sentences of the same type, introducing noise. To address this, we use LLMs as a neat tool to rewrite sentences in a more standardized manner, enabling more uniform syntactic graphs and reducing noise caused by diverse writing styles.

In this paper, we propose the SAM4LTC (**S**yntax **A**ware **M**asked model for **L**ife **T**rajectory **C**lassification) framework to fuse syntactic information into text representations for fine-grained classification of life trajectory activities. Its input is the life trajectory triple and its context sentence extracted from Wikipedia by COSMOS [39] and the syntactic structure of the sentence is refined using LLM. It first constructs the syntactic graph of the life trajectory activity sentence and then extracts the subgraph of the words most related to the triple by searching the shortest paths among the triple elements. Finally, additional supervised contrastive learning loss is adopted to enhance the model's ability to learn to distinguish different classes. To train SAM4LTC, we manually annotate 2,826 life trajectory activities randomly selected from Wikipedia bio pages. We summarize the contributions of our work as follows:

- We propose a new life trajectory classi task and a taxonomy of life trajectory activities, as well as a manually annotated dataset on this taxonomy.
- The SAM4LTC framework fuses syntactic and semantic information achieving an Accuracy of 84.5%. Additionally, we refine syntactic structures with LLM to improve data quality.
- We release our model, manually-annotated dataset, and the largest fine-grained life trajectory dataset to date (3.8 million tuples)[1] to support research on human dynamics and information extraction. Analysis of 589,193 individuals across 3 centuries demonstrates how fine-grained trajectories enable insights into human dynamics.

2 Related Work

2.1 Life Trajectories and Human Dynamics

Life trajectory analysis informs human dynamics. Studies of over 300,000 births and deaths trace cultural evolution through migration [24], and artist migration is linked to the rise and fall of creativity centers [25]. With LLM-based simulations, trajectory data gains value beyond analytics [26,21]. Historical profiles (birth, death, official posts) have been used to fine-tune LLMs for interactive role-playing [4]. Large-scale extraction efforts address data scarcity by producing high-quality triples from Wikipedia [39]. Multiple life trajectory activities are needed to build a complete life course.

[1] `https://anonymous.4open.science/r/SAM4LTC-B7C9/` We release 247,292 tuples in this version; all 3.8M tuples will be released upon acceptance.

2.2 Classification with Syntactic Graphs

Traditional sentence classification relies on pre-trained language models (PLMs) [6,35] for holistic embeddings, while graph-based methods exploit syntactic structures [36,15]. Aspect-level sentiment classification (ALSC) similarly employs syntactic graphs to connect aspects and opinions [13,12,30]. In our task, life trajectory triple entities are close in the syntactic graph but distant in the raw sentence. However, existing approaches do not tightly fuse structure information with text embeddings throughout training. To achieve deeper fusion, we introduce a MASK generation method built upon the syntactic graph, which directs the model to focus on the target life trajectory triple among multiple entities.

3 Problem Statement and Dataset

3.1 Problem Statement

Given a Wikipedia bio page, we use COSMOS to extract the sentence S, and the triple $T = (person, time, location)$, where the elements in the triple T are in the sentence S, the sentence describes the life trajectory activity that the $person$ is experiencing his/her life event in the $location$ at the $time$. In addition, the label space $L = (l_1, l_2, \cdots, l_n)$ represents the taxonomy (will be introduced in Dataset Sec.) of the life trajectory activities. We train a model f to assign a label (i.e., the type of the life trajectory activity that feeds into f) in $l \in L$ to each triple T based on the sentence S.

3.2 Data Collection, Taxonomy and Annotation of Life Trajectories

The data comes from English Wikipedia biography pages. We first selected candidates via Wikidata entities labeled *instance of human*, then randomly chose 2,000 human biography pages. Using the COSMOS tool, we extracted life trajectory triples and retained their source sentences.

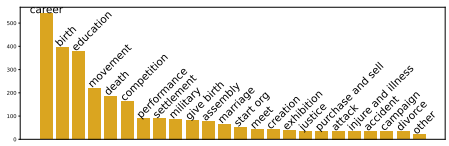

Fig. 2: Distribution of types in the *Regular* dataset.

As mentioned in Introduction, existing research mainly analyzes births and deaths, and has not yet specifically designed a taxonomy for life trajectory activities. Drawing on the types of events in the task of event extraction ACE [7], studies in related fields [10,11], and our empirical observation of the data, we propose the following taxonomy of life trajectory activities, consisting of 24 types in 9 categories. Details of the taxonomy can be found in the code repo. We note that this taxonomy may be to some extent arbitrary, and further research may need to adjust it according to specific needs. Based on the taxonomy, we manually annotate a dataset of 2,826 samples from the extracted life trajectory triples for training and testing our model. In addition, the distribution of our dataset is shown in Fig. 2. We also validate the annotating consistency and the Cohen's kappa between two annotators is 0.97, demonstrating that our taxonomy is reproducible.

3.3 Syntax Refinement using LLM

Previous work use LLM to annotate data [32,19] or directly generate data samples satisfying their requirements [16,34] to augment training data for better performance of their models. Unlike these studies, we do not seek to generate new data, but to refine the syntactic structures of the sentences by requiring LLM to rewrite the sentences in another way without changing the original meaning. This is because we are concerned that training on samples generated by LLM may lead to overfitting issues, making it difficult to generalize on real-world samples. Simply rewriting the sentences preserves real-world data distribution while providing a more standardized syntactic structure for the model to learn. Specifically, we require the LLM to rewrite the sentence S to get the refined sentence S_{ref}. The prompts we use for refinement can be found in the code repo.

Table 1: Taxonomy of Life Trajectory Activities

Category	Type
Life	Birth, Death, Marriage, Injure and Illness, Settlement, Education, Give birth, Accident, Purchase and Sell, Divorce
Career	Career
Pro-Event	Competition, Performance, Exhibition, Creation, Campaign, Start org
Contact	Meet, Assembly
Justice	Justice
Attack	Attack
Movement	Movement
Military	Military
Other	Other

We refer to the fully manually annotated dataset as the *regular* dataset, while the dataset refined by the LLM is called the *LLM-Refined* dataset. Notably, when LLM occasionally fails to rewrite a sentence, we implement a simple quality control mechanism by prompting it to redo the task until the sentence has been successfully modified.

4 Method

4.1 Overview of the Method

As shown in Fig. 3, the SAM4LTC workflow first employs an LLM to refine the input sentence S without altering any words from the triple T or changing the sentence meaning, while inserting special tokens around the triple elements so that they are perceived as holistic units. From S and T, a graph-based trajectory subgraph is then constructed, where nodes correspond to tokens produced by a PLM tokenizer and edges represent dependency relations parsed by SpaCy. Based on this trajectory subgraph, a 'MASK' vector is derived to obtain embeddings of the words most relevant to the trajectory activity, which are concatenated with the sentence embedding from ERNIE [41]; finally, a fully connected layer yields the classification. ERNIE is adopted as the preferred PLM because, as a top-ranked open-source model on the GLUE [29] benchmark, its pretraining strongly emphasizes entity information, thereby facilitating the understanding of person, time, and location entities in the triple.

4.2 Graph Construction

For the input sentence $S = \{w^1, w^2, ..., w^m\}$ with length m and triple $T = \{t_p, t_t, t_l\}$ extracted from S, we first cut them into subwords with ERNIE tokenizer, which become $S = \{w^1, w^{1\#1}, \cdots, w^m, \cdots, w^{m\#k}\}$ and $T = \{t_p^1, \cdots, t_p^{p\#k}, t_l^1, \cdots, t_l^{l\#k}\}$, where $w^{m\#k}$ is the k-th subword of w^m and t_p, t_t, t_l is the words indicating person, time and location with word length of p, t, l. Then we add special tokens to the vocabulary of the tokenizer to mark the corresponding triple elements in the sentence S.

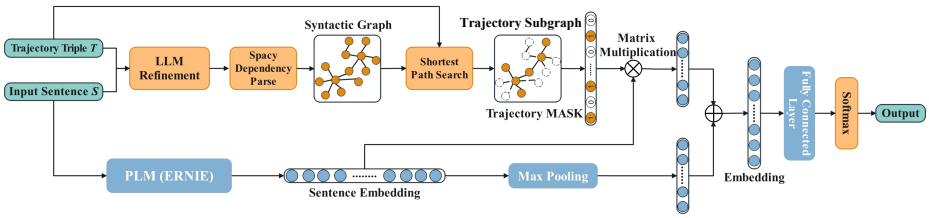

Fig. 3: Workflow of SAM4LTC. The trainable modules and embeddings are in blue.

Upon tokenizing the given sentence and words in triple into tokens, we utilize SpaCy to construct a graph $G = (V, E)$ of the sentence based on the syntactic structure and dependency relations within the sentence parsed by SpaCy. In this overall graph, nodes correspond to the tokens in S, and edges denote their dependencies. Additionally, if words were tokenized by the ERNIE tokenizer, we will add edge between them.

Verbs typically serve as the bridge to connect the entities in a sentence. Hence, we search the shortest paths between the words in T and their nearest verbs and use the words on the paths to extract a local subgraph from the overall graph to represent the connection among the trajectory triple. Specifically, let SP be the shortest path algorithm and $Verb$ be the verbs in the sentence, then compute 3 paths:

$$Path(t) = min(SP(t, Verbs)), \tag{1}$$

$$Path = Path(t_p) \cup Path(t_t) \cup Path(t_l), \tag{2}$$

where $SP(A, B)$ returns all the shortest paths between nodes in A and nodes in B, t_p, t_t, t_l are the nodes corresponding to person, time and location in T.

Then the local subgraph $G' = (V', E')$ can be extracted from the overall graph using the nodes in $Path_p$, $Path_t$ and $Path_l$:

$$V' = Path, \tag{3}$$

$$E' = \{(u, v)\},\ u, v \in V', \tag{4}$$

Note that each sentence extracted from Wikipedia consist limited words and therefore the additionally introduced computation is trivial comparing to the neural network though the $O(n^3)$ shortest path algorithm is adopted in the framework.

4.3 Representation with Generated MASK

After obtaining the local subgraph, we compute the 'MASK' vector by simply marking the corresponding position as 1 if the corresponding subword is in G or V' and 0 otherwise:

$$MASK[i] = \begin{cases} 1 & \text{if } i\text{-th token of } S \in V', \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

To get the representation of the trajectory activity, we perform matrix multiplication on the 'MASK' vector and the embeddings output by ERNIE h_e. The trajectory activity embedding h_1 is obtained by masking the $MASK$ on h_e, which is $h_1 = MM(h_e, MASK)$, where MM is the matrix multiplication. Then mean-pooling is adopted, which is $h_1' = MeanPooling(h_1)$. The contextual embeddings h_2 are obtained the through max-pooling of h_e, $h_2 = MaxPooling(h_e)$, where $MaxPooling$ is max-pooling operation. Finally, the embedding used for contrastive learning and cross entropy loss is obtained by concatenating h_1 and h_2:

$$h_{scl} = h_1 \oplus h_2, \tag{6}$$

4.4 Contrastive Learning and Cross-Entropy Loss

Conventional cross-entropy loss is inadequate in our scenario of insufficient annotated data [42], for the possibility of overfitting and poor generalization. Supervised contrastive learning (SCL) integrates the idea of "learn to compare" into supervised learning [14], mitigating the issue by reducing reliance on labels and improving the model's generalization ability.

For the output feature vector h_{scl}, we compute the supervised contrastive learning loss as follow:

$$L_{scl} = \sum_{i=1}^{N} \frac{-1}{N_{y_i} - 1} \sum_{j=1}^{N} \mathbf{1}_{i \neq j} \mathbf{1}_{y_i \neq y_j} log \frac{e_{ij}}{\sum_{k=1}^{N} \mathbf{1}_{i \neq k} e_{ik}},$$

where the $e_{mn} = exp(h_{scl}^m \cdot h_{scl}^n / \tau)$ and the cross-entropy loss is computed by $L_{ce} = -\frac{1}{N} \sum_i^N (y_i log(y_i') + (1 - y_i) log(1 - y_i'))$, where N is the batch size, h_{scl} is the embedding from the model, N_{y_i} is the total number of examples with the same labels as y_i in a batch and τ is the temperature parameter controlling the dissimilarity of samples from different classes should be. In cross-entropy loss, y_i' is the predicted label by the model. Finally, the loss function used to train our model is combined with L_{scl} and L_{ce} adjusted by a hyperparameter λ:

$$L = (1 - \lambda) L_{ce} + \lambda L_{scl}. \tag{7}$$

5 Experiments

5.1 Experimental Setup and Baselines

Experimental Setup. We perform 10-fold cross-validation and use Precision, Recall, and F1-score as metrics. They are weightedly calculated according to the number of

samples in each class, which makes Recall mathematically equal to Accuracy. Results with the best Recall among the training epochs are reported.

Baselines. We compare against 8 baselines of different specialties. TextGCN [36], BERT [6], XLNet [35] and ERNIE [41], are graph-based or PLMs having shown good performance on text classification. R-GAT [30] is a graph-based model developed for ALSC, sharing certain commonalities with our task and model in terms of task definition and information utilization. GPT-4 [1] and GPT-5 are also included as the baseline. EvoPrompt [9] is a prompt learning method that optimizes the discrete prompt via evolution algorithms to fit specific tasks. The prompts can be found in the code repo.

5.2 Experimental Results

Prediction Performance As reported in Table 2, our model outperforms the baselines across all experiments. When transitioning from the *Regular* dataset to the *LLM-Refined* dataset, SAM4LTC consistently outperforms ERNIE in all 10 folds, achieving mean F1 scores of 84.5% vs. 82.9%. The mean improvement is 1.53% (±0.79% std), which is statistically significant ($p \leq 0.0005$). We further recalculate the metrics by categories (9 in total, see Table 1) instead of the fine-grained types (24 in total), all methods naturally reach higher F1-scores due to this relaxation, and ours hits the F1-score of 89.6%.

Table 2: Performance comparisons for life trajectory activity classifications.

Model	Regular/Type			LLM-Refined/Type			LLM-Refined/Category		
	P	R (Acc)	F1	P	R (Acc)	F1	P	R (Acc)	F1
TextGCN	62.1%	61.3%	29.3%	63.4%	61.5%	60.3%	76.8%	71.7%	71.7%
R-GAT	74.1%	78.0%	74.8%	71.3%	75.8%	72.2%	81.4%	81.2%	80.7%
BERT	82.5%	81.6%	81.4%	83.9%	83.0%	82.8%	84.7%	84.7%	84.6%
XLNet	82.5%	81.7%	81.4%	84.2%	83.3%	83.0%	85.6%	85.8%	85.7%
ERNIE	83.1%	82.1%	81.8%	84.2%	83.2%	82.9%	85.7%	85.9%	85.8%
GPT-4	73.9%	76.2%	72.6%	75.1%	72.6%	70.0%	80.6%	81.5%	80.5%
GPT-5	71.2%	73.5%	70.0%	74.7%	74.6%	72.9%	78.8%	79.6%	78.4%
EvoPrompt	74.2%	71.6%	70.8%	76.1%	73.9%	74.3%	80.6%	81.5%	80.5%
SAM4LTC	**83.7%**	**82.5%**	**82.2%**	**85.2%**	**84.5%**	**84.5%**	**86.8%**	**87.0%**	**86.9%**

Among the baselines, TextGCN tends to underperform compared to most other methods. This may be attributed to the sparsity of the adjacency matrices constructed from word and sentence co-occurrences, as well as the oversight of triple word relationships, both of which could negatively impact its performance. R-GAT, which is intended for the ALSC task, achieves an F1 score of only 72.2% for type classification on *LLM-Refined*, falling short of our expectations. A possible reason for this may be that R-GAT is designed to focus on a single aspect entity and its corresponding sentiment, while the triples of three entities in our scenario may exceed its capacity. GPT-4, GPT-5 and EvoPrompt achieve Recall under 80.0%, suggesting prompt-based methods may struggle with this task's requirement for precise triple-context alignment. Even EvoPrompt's evolutionary prompt optimization cannot close this gap, likely because the scattered triple entities and complex syntactic relationships are difficult to capture

through prompting alone. The 7 least frequent classes achieve 79.9%, demonstrating strong generalization across all life event types.

5.3 Ablation Study

For the ablation study, we remove the syntactic MASK vector and the SCL to show their necessity. Meanwhile, we check the necessity of classifying the triples along with their contextual sentences, instead of classifying the sentences only, by removing the components for capturing the syntactic graphs and only keeping the ones generating sentence embedding in Fig. 3. The results are shown in Table 3, none of the ablation experiments exceed the SAM4LTC.

Table 3: Model ablation results.

Model	Pre (%)	R (Acc) (%)	F1 (%)
$w/o\ ma$	82.2 (-3.0)	82.9 (-1.6)	82.2 (-2.3)
$w/o\ scl$	83.3 (-1.9)	82.6 (-1.9)	82.3 (-2.2)
$w/o\ tri$	83.2 (-2.0)	82.5 (-2.0)	82.2 (-2.1)

Compared to SAM4LTC, SAM4LTC$_{w/o\ ma}$, which uses MASK that only masks the elements in the triple instead of the MASK from the local subgraph, decreases in F1 by 2.3%. Secondly, SAM4LTC exceeds SAM4LTC$_{w/o\ scl}$ that removes the supervised contrastive loss and only leaves cross-entropy loss by 2.2% in F1. Finally, the SAM4LTC$_{w/o\ tri}$ without triples and the relevant syntactic structures falls behind SAM4LTC by 2.1% in F1.

6 Analysis of the Extracted Dataset

We validate our method and analyze labeled life trajectory activities to show the potential of the method as well as the data. We classify all trajectory triples extracted from the English Wikipedia, followed by coreference resolution and geocoding.

In all, we have a *3-century* dataset of 3,828,133 triples of 589,193 people to start with. For a closer look, we narrow it down to 20th-century Germany (*20th-Germany*) and the United States (*20th-US*), respectively including 17,698 triples of 5,435 Germans and 229,594 triples of 75,149 Americans.

Data at a Glance. From the *3-century* dataset, we randomly sample 207 triples and find that 80.16% are accurate. The type distribution of the *3-century* dataset is similar to Fig. 2. Fig. 4 (a) showcases the life trajectory of American singer. *Malcolm John Rebennack Jr.* who ranks among top 20% in the dataset, according to number of triples. Born in New Orleans, he lived in Los Angeles in the West, created his works in Baltimore and spent his main performing career in the East in the 70s and 80s. In his later years, he returned to his hometown and died in 2019.

Wars and Sports in Three Centuries. We examine two activity types, *military* and *competition*. Within each 5-year range, we calculate the percentage of the specific type of activity among all activities and the overall distributions of these two activities are visualized in Fig. 4 (b). The peaks in the curve for *military* correspond well with major wars temporally, from the Spanish Succession War to WWII. *Competition*, on the other hand, witnesses a relatively steady growth, starting from the late 19th century,

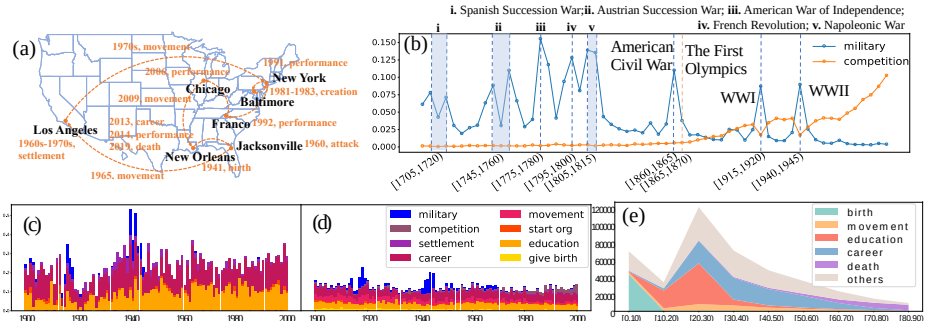

Fig. 4: (a) Life trajectory of Malcolm John Rebennack Jr. (b) Ratios of military and competition activities from 1700 to 2000 in five-year intervals. Wars and the First Olympics are marked. (c) and (d) International departures from Germany and the US in the 20th century. (e) Activities in life stages in a stacked chart. X-axis marks the age group and y-axis is the number of life trajectory activities.

consistent with the expansion of popular sports, marked by the first modern Olympics in 1896. Curiously, the *competition* trend obviously drops during WWI and WWII, and it actually negatively correlates with the *military* trend during the 20th century (Pearson correlation coefficient being -0.52, with a p-value of $2.42e^{-5}$).

How/Why People Leave Home Countries. We zoom in and conduct fine-grained analysis on Germany and the United States to examine how and why people depart their home countries. We select the top 8 types except birth and death for the analysis. An international departure is a move from the home country to a foreign one, judged by two consecutive trajectory triples of a same person. Fig. 4 (c) and (d) visualize the ratios of international departures against all travels started within the home country (including international departures and the travels within the home country), for the two countries respectively. For Germany, we see a surge in international departures around the 1930s and early 1940s which may be associated with the rule of Nazi. If we ignore the *military* activities in blue, the surge still persists. Its larger contributions are from the increase in *career* and *settlement*, which may indicate a brain drain. In fact, during this period, many Jewish figures, such as Albert Einstein and Hannah Arendt (philosopher) fled Germany. If we look at the US, its ratio of international departures is relatively lower. Its two surges correspond to the two world wars. Interestingly, *military* activities mainly contribute to these two peaks and the US does not seem to suffer from brain drain.

Individual Life Stages and Activities. We further focus on the Americans in the 20th century and examine their life stages and activities aggregated at the individual level. We visualize different trajectory activities throughout different life stages, with a stacked chart aggregating these people's activities (Fig. 4 (e)). In each age group, we show the total amounts for the *birth*, *death*, *education*, *career*, and *movement* types of activities, which account for about 70% samples of the dataset. Rest activities are labeled as *others*. It shows that people in their 20s have the most trajectory activities, with the top ones being *education* and *career*, while the death type takes up larger portions beyond age 70.

7 Conclusion

We propose a new task of labeling life trajectory activities of people and introduce the taxonomy including 24 types in 9 categories of life trajectory activities with a manually annotated dataset. We design SAM4LTC fusing syntactic structure into text representation to classify life trajectory activities and use LLM to refine the syntactic structures of sentences. Our model outperforms the baselines in the experiments. Besides, we conduct an empirical analysis using the life trajectory activities with corresponding types.

References

1. Achiam, J., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Aleksanyan, M., Hao, Z., Vagenas-Nanos, E., Verwijmeren, P.: Do state visits affect cross-border mergers and acquisitions? Journal of Corporate Finance **66**, 101800 (2021)
3. Atari, M., Henrich, J., Schulz, J.: The chronospatial revolution in psychology. Nature Human Behaviour pp. 1–9 (2025)
4. Bai, T., Kang, J., Fan, J.: Baijia: A large-scale role-playing agent corpus of chinese historical characters. arXiv preprint arXiv:2412.20024 (2024)
5. Dai, J., Yan, H., Sun, T., Liu, P., Qiu, X.: Does syntax matter? a strong baseline for aspect-based sentiment analysis with roberta. arXiv preprint arXiv:2104.04986 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S.M., Weischedel, R.M.: The automatic content extraction (ace) program-tasks, data, and evaluation. In: LREC. vol. 2, pp. 837–840. Lisbon (2004)
8. Elder Jr, G.H.: Time, human agency, and social change: Perspectives on the life course. Social psychology quarterly pp. 4–15 (1994)
9. Guo, Q., et al.: Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. In: ICLR (2024)
10. Herrera-Guzmán, et al.: Quantifying hierarchy and prestige in us ballet academies as social predictors of career success. Scientific Reports **13**(1), 18594 (2023)
11. Huang, J., Gates, A.J., Sinatra, R., Barabási, A.L.: Historical comparison of gender inequality in scientific careers across countries and disciplines. PNAS **117**(9), 4609–4616 (2020)
12. Huang, L., Sun, X., Li, S., Zhang, L., Wang, H.: Syntax-aware graph attention network for aspect-level sentiment classification. In: COLING. pp. 799–810 (2020)
13. Ke, W., Gao, J., Shen, H., Cheng, X.: Incorporating explicit syntactic dependency for aspect level sentiment classification. Neurocomputing **456**, 394–406 (2021)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS **33**, 18661–18673 (2020)
15. Lai, V.D., Nguyen, T.N., Nguyen, T.H.: Event detection: Gate diversity and syntactic importance scores for graph convolution neural networks. In: EMNLP. pp. 5405–5411 (2020)
16. Li, Z., Zhu, H., Lu, Z., Yin, M.: Synthetic data generation with large language models for text classification: Potential and limitations. arXiv preprint arXiv:2310.07849 (2023)
17. Liu, Y., Medlar, A., Glowacka, D.: Can language models identify wikipedia articles with readability and style issues? In: SIGIR. pp. 113–117 (2021)
18. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J.: Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR) **54**(3), 1–40 (2021)

19. Møller, A.G., Pera, A., Dalsgaard, J., Aiello, L.: The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In: EACL. pp. 179–192 (2024)
20. Park, J., Wood, I.B., Jing, E., Nematzadeh, A., Ghosh, S., Conover, M.D., Ahn, Y.Y.: Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. Nature communications **10**(1), 3449 (2019)
21. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: UIST. pp. 1–22 (2023)
22. Qian, S., Chen, H., Xue, D., Fang, Q., Xu, C.: Open-world social event classification. In: Proceedings of the ACM Web Conference 2023. pp. 1562–1571 (2023)
23. Quarles, C.L., Budak, C., Resnick, P.: The shape of educational inequality. Science Advances **6**(29), eaaz5954 (2020)
24. Schich, M., Song, C., Ahn, Y.Y., Mirsky, A., Martino, M., Barabási, A.L., Helbing, D.: A network framework of cultural history. Science **345**(6196), 558–562 (2014)
25. Serafinelli, M., Tabellini, G.: Creativity over time and space: A historical analysis of european cities. Journal of Economic Growth **27**(1), 1–43 (2022)
26. Shanahan, M., McDonell, K., Reynolds, L.: Role play with large language models. Nature **623**(7987), 493–498 (2023)
27. Vaswani, A., et al.: Attention is all you need. NeurIPS **30** (2017)
28. Verginer, L., Riccaboni, M.: Cities and countries in the global scientist mobility network. Applied Network Science **5**, 1–16 (2020)
29. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R.: Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
30. Wang, K., Shen, W., Yang, Y., Quan, X., Wang, R.: Relational graph attention network for aspect-based sentiment analysis. arXiv preprint arXiv:2004.12362 (2020)
31. Wang, Y., Jones, B.F., Wang, D.: Early-career setback and future career impact. Nature communications **10**(1), 4331 (2019)
32. Whitehouse, C., Choudhury, M., Aji, A.F.: Llm-powered data augmentation for enhanced cross-lingual performance. In: EMNLP (2023)
33. Wolfinger, N.H., Mason, M.A., Goulden, M.: Problems in the pipeline: Gender, marriage, and fertility in the ivory tower. The Journal of Higher Education **79**(4), 388–405 (2008)
34. Xu, F., other: Towards large reasoning models: A survey of reinforced reasoning with large language models (2025)
35. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. NeurIPS **32** (2019)
36. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: AAAI. vol. 33, pp. 7370–7377 (2019)
37. Yen, A.Z., Huang, H.H., Chen, H.H.: Personal knowledge base construction from text-based lifelogs. In: SIGIR. pp. 185–194 (2019)
38. Zhang, M., Qian, T.: Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In: EMNLP. pp. 3540–3549 (2020)
39. Zhang, Y., Li, X., Liu, Z., Zhang, H.: Paths of a million people: Extracting life trajectories from wikipedia. In: ICWSM. vol. 19, pp. 2226–2240 (2025)
40. Zhang, Y., Melnykov, V., Zhu, X.: Model-based clustering of time-dependent categorical sequences with application to the analysis of major life event patterns. Statistical Analysis and Data Mining: The ASA Data Science Journal **14**(3), 230–240 (2021)
41. Zhang, Z., et al.: Ernie: Enhanced language representation with informative entities. In: ACL. pp. 1441–1451 (2019)
42. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. NeurIPS **31** (2018)