歴史テキストからの動的知識グラフ構築と探索的分析手法の提案

戸田 隆道 [1] 平野 智也 [1]
[1] 株式会社 COTEN
{takamichi.toda, tomoya}@coten.co.jp

概要

歴史学習や研究において，個別の出来事の記憶以上に，国家や都市間の協力・対立といった「関係構造」の変遷を捉えることが重要である．しかし，教科書等の記述は列挙的であり，学習者が全体構造を俯瞰することは困難であった．解決策として人手による関係アノテーションが考えられるが，コストが高く，大規模データへの適用が難しい．本研究では，書籍内での固有表現の共起を「関係の存在」の近似とみなし，LLM により各エンティティの時代・地理位置・関係極性を自動付与することで，精度を一定程度犠牲にしつつも全体構造のインサイトを得られる探索的分析手法を提案する．国名・都市名に対象を限定し，抽出した関係を地図上に時系列で可視化する Web アプリケーションを構築した．

1 はじめに

歴史学習・歴史研究においては，個別の出来事の知識だけでなく，出来事間の関係（誰が誰と協調し，どこで対立し，その構造がいつ変わったか）を捉えることが重要である．しかし，教科書や事典などの記述から抽出されるイベントデータは，出来事が列挙的に並ぶ傾向があり，出来事間の結びつきが疎（スパース）になりやすい．そのため，テキストから全体構造を直接理解することは難しい．

この課題に対し，出来事間の関係を一つ一つ紐付け，ラベリングしていくアプローチが考えられる．だが，こうした手作業には膨大なコストがかかり，大規模なテキストへの適用は現実的ではない．そこで本研究では，精度を一定程度犠牲にしても，統計的・自動的に関係をラベリングする手法を目指す．

本研究の基本的なアイデアは以下の通りである：

1. 共起による関係の近似：書籍の複数ページにわたって共起する固有表現は，何らかの関係を持つと仮定する
2. 対象の限定：まずは国名・都市名に対象を絞り，地政学的な関係構造の分析に焦点を当てる
3. LLM による属性付与：各固有表現について，関連する年代・地図上の位置・関係の極性（Positive/Negative/Neutral）を LLM に自動抽出させる
4. 全体インサイトの重視：個々の抽出結果には不正確さが含まれうるが，全体を俯瞰した際に得られるインサイトを重視する

関係の極性を付与することで，単に「関係がある」という情報を超えて，協力関係か対立関係かを区別し，さらにその関係が時間とともにどう変化したかを追跡することが可能となる．

この手法に基づき，抽出した関係を地図上に時系列で可視化する Web アプリケーションを構築した．本稿では，提案手法の設計と実装，および世界史教科書を対象とした分析事例を報告する．

2 関連研究

歴史テキストを対象とした情報抽出として，固有表現認識（NER）やイベント抽出の研究が進められてきた．Sprugnoli ら [3] は歴史文書に特化したイベント検出・分類のガイドラインを提案し，歴史叙述に現れるイベントカテゴリを整理している．また Lai ら [1] は 19 世紀の新聞記事を対象にイベント抽出データセットを構築し，既存モデルの適用可能性を検証した．一方で，抽出された関係が協調か対立かといった関係の意味的側面（極性）については十分に扱われていない．

計算歴史学（Computational History）の分野では，大規模なデジタル化された歴史資料を統計的手法で分析し，過去と現在を結びつける研究が進められている [6]．これらの研究は，個別の出来事の抽出を超えて，歴史的パターンの発見や長期的な変化の可

視化を目指しており，本研究の目的と方向性が一致する．

時間情報を組み込んだ知識グラフ（Temporal Knowledge Graph; TKG）の研究も進展している．Trivedi ら [4] の Know-Evolve に代表されるように，時間とともに変化する表現を学習し，リンク予測や事象発生時刻の予測を目的とする枠組みが提案されてきた．しかし，これらの研究は主に「未来のリンクを当てる」ためのモデル化が中心であり，歴史的記述の理解を支援するために，過去の関係構造を直観的に俯瞰する可視化や探索的分析の検討は限定的である．

近年は大規模言語モデル（LLM）を関係抽出に利用する試みが増えており，Few-shot プロンプトによる柔軟な関係分類が有効であることも示されている [5]．本研究はこの流れを歴史テキストへ応用し，共起で得た関係候補に対して LLM により**時間・地理・関係極性**を自動付与することで，精度よりも網羅的な構造把握を優先した動的知識グラフの構築と可視化による探索的分析を提案する点に特徴がある．

歴史データの可視化に関しては，Michel ら [2] が数百万冊のデジタル化された書籍を用いて文化的変化を時系列で可視化する手法を提案しており，大規模な歴史データの可視化によるマクロなパターン発見の有効性を示している．本研究は，このような可視化による探索的分析のアプローチを，関係構造の時空間的表現に特化して適用する点に特徴がある．

3　提案手法

3.1　全体設計

提案手法は個々の抽出精度を追求するのではなく，大量の関係を自動処理し，集約・可視化することで全体構造のパターンを浮かび上がらせることを目指す．また，本研究でいう動的知識グラフは，(1) 国・都市をノード，(2) 共起に基づく関係候補をエッジ，(3) エッジに時間区間と極性ラベルを付与し，時間に応じてサブグラフが変化する表現を指す．すなわち，本研究はリンク予測を目的とする Temporal Knowledge Graph（TKG）とは異なり，歴史理解のための探索的可視化・分析を主目的とする．

3.2　共起に基づく関係抽出

3.2.1　固有表現の抽出と正規化

入力テキストから固有表現を抽出し，以下のフィルタリングを行う：

- **対象エンティティ**：国名，都市名に限定
- **正規化**：表記ゆれの統一（例：「米国」「アメリカ」「アメリカ合衆国」→「アメリカ」）

国名・都市名への限定は，地理的可視化との親和性が高いことに加え，地政学的構造の分析に直結する，という理由による．

3.2.2　共起関係の定義

エンティティペア e_i, e_j が同一ページ内に出現するとき，共起関係があるとみなす．共起頻度 $\mathrm{freq}(e_i, e_j)$ を重みとしてエッジを構築し，閾値以上のペアのみを関係候補として抽出する．

3.3　LLM による属性の自動付与

抽出された各エンティティおよび関係に対し，LLM を用いて以下の 3 属性を付与する．

- **時代**：エンティティが文脈中で言及される年代・時期を抽出する．
- **地理位置**：国名・都市名から緯度・経度を LLM で推定する．範囲が広大な場合はその中央とする．
- **関係極性**：各エンティティが文脈中でどのような立場や評価で言及されているかを，Positive / Negative / Neutral の 3 クラスに分類する．

4　可視化による探索的分析

4.1　システム構成

本章では，提案手法により構築された動的知識グラフを用いて，歴史テキストに内在する関係構造を探索的に分析するための可視化手法と，その分析可能性について述べる．

本研究における可視化は，結果を単に提示するためのデモではなく，不完全かつノイズを含む自動抽出結果から，マクロな構造的インサイトを得るための分析インターフェースとして位置づけられる．

4.2　探索的分析タスクの定義

歴史テキストから自動構築された関係ネットワークに対して，以下の探索的分析タスクを想定する．

T1: 地域間関係の空間的構造

ある年代・時期において，どの地域・国家間に協力関係や対立関係が集中しているかを俯瞰的に把握する．

T2: 関係構造の時間的変化

同盟や対立の形成・解消といった関係極性の変化が，時間軸上でどのように推移するかを観察する．

T3: ネットワーク内の中心エンティティ

特定時代において，多数の関係を持つ国家・都市や，対立・協力のハブとなるエンティティを同定する．

これらのタスクは，個々の関係の正確性よりも，集約された関係構造の傾向や変化点を捉えることを目的とする点に特徴がある．

4.3　可視化設計

上記タスクを支援するため，時空間情報と関係極性を統合的に扱うインタラクティブな可視化を設計・実装した．

4.3.1　時空間マッピング

地図上に国・都市をノードとして配置し，関係をエッジとして描画する．エッジの色は極性に対応する：

- 青：Positive（協力関係）
- 赤：Negative（対立関係）
- 灰：Neutral（中立的関係）

T1（地域間関係の空間的構造）を直感的に支援する．

4.3.2　時代スライダー

関係に付与された時代情報に基づき，時代スライダーを用いて表示対象の年代を制御できるようにした．これにより，特定時期の関係構造を切り出して表示したり，連続的な時間変化を追跡したりすることが可能である．この操作により，例えば戦争前後や政体変化前後における関係構造の変化を視覚的に比較でき，T2（関係構造の時間的変化）を支援する．

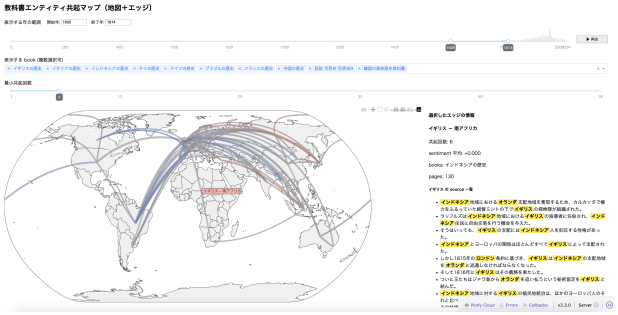

図 1　可視化システムの画面例．1790 年代のヨーロッパにおける関係構造を表示．青線は協力関係，赤線は対立関係を示す．

4.3.3　詳細情報と原文参照

ノードやエッジを選択すると，対応する元テキストの該当箇所を表示する詳細パネルを提供する．これにより，ユーザーは自動抽出された関係の根拠となる記述を確認できる．この設計は，LLM による自動付与結果をブラックボックス化せず，解釈可能性と検証可能性を確保することを目的としている．また，誤抽出を含みうる結果を前提とした探索的分析において，ユーザーが必要に応じて判断を下せる余地を残す点で重要である．

4.4　分析事例

本研究では，提案手法の挙動を確認するための一例として，10 冊の歴史の教科書を対象に関係ネットワークを構築し，その時間的推移を可視化した．

4.4.1　実験設定

検証では明石書店の「世界の書籍シリーズ」から「イタリアの歴史」，「フランスの歴史」，「タイの歴史」，「イギリスの歴史」，「中国の歴史」，「インドネシアの歴史」，「ドイツの歴史」，「ブラジルの歴史」，「韓国の高校歴史教科書」，及び山川出版社「詳説世界史 B」を対象テキストとして使用した．

LLM には Gemini 2.5 Flash Lite を用い，各属性付与は Few-shot プロンプトで実行した．抽出の結果，固有表現として国名 130 件，都市名 115 件（正規化後）を得た．共起ペアは閾値適用後に 9882 件となり，うち Positive 549 件（5.6%），Negative 2518 件（25.4%），Neutral 6815 件（69.0%）に分類された．

4.4.2　可視化結果

構築した Web アプリケーションにより，国家・都市間の関係を時系列で可視化した結果（図 1），時代の進行に伴って国際関係ネットワークの中心構造が

段階的に変化する様子が確認できた．1500–1799 年は西欧を核に遠隔地（南北アメリカやアジア・オセアニア）へ伸びる長距離の結び付きが目立ち，海外領域をめぐる関係がネットワークの骨格を形づくる．19 世紀になると，欧州を中心としたまま列強間の相互関係が強まり，ユーラシア方向を含む大陸規模の結び付きが太く現れる．1900–1945 年は世界大戦期を含むため、欧州周辺の主要国間の結び付きがコアとして凝集し．欧州—北米・欧州—ロシアを結ぶ幹線が際立つ．戦後の 1946–1969 年には．米国とロシア（ソ連）が双ハブとして前面に現れ．米露・米欧・露欧といった大国間関係が太く可視化されると同時に．旧宗主国—独立側の関係も上位に現れ．二極構造と脱植民地化に伴う再編が同じ図の中に重なって見える．

このように，本可視化は個別の出来事を逐次的に追跡することなく，教科書叙述に内在する国際関係のマクロな変遷を俯瞰的に把握する手段として機能することが示された．なお，T3（中心性分析）の定量化については，複数教科書への適用等の拡張と合わせ，今後の課題とする．

5 おわりに

本研究では，歴史テキストから共起に基づいて関係を抽出し，LLM で時代・位置・極性を自動付与することで，動的知識グラフを構築する手法を提案した．本手法は個々の抽出精度を犠牲にする代わりに，手作業では困難な規模の関係ネットワークを構築・可視化し，歴史の「大きな流れ」を俯瞰することを可能にした．

本論文の主な貢献として，歴史テキストから自動構築された関係ネットワークに対する探索的分析タスク（空間的構造の把握，時間的変化の観察，中心エンティティの同定）を明確化し，それぞれに対応する可視化設計を提案した点が挙げられる．

限界として，対象エンティティを国名・都市名に限定している点，関係の意味的分類を三値に単純化している点，共起に基づく相関的な関係であり因果関係を表すものではない点がある．今後は，人物・組織への対象拡張，関係タイプの細分化，および複数教科書間での再現性検証を進め，歴史学習・研究支援ツールとしての実用性を高めていきたい．

参考文献

[1] Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. Event extraction from historical texts: A new dataset for black rebellions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pages 2390–2400, Online, August 2021. Association for Computational Linguistics.

[2] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. **Science**, 331(6014):176–182, 2011.

[3] Rachele Sprugnoli and Sara Tonelli. Novel event detection and classification for historical texts. **Computational Linguistics**, 45(2):229–265, June 2019.

[4] Rakshit Trivedi, Hanjun Dai, Yichen Wang, and Le Song. Know-evolve: Deep temporal reasoning for dynamic knowledge graphs, 2017.

[5] Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large language models, 2024.

[6] アダム ヤトフト．計算歴史学による過去と現在の橋渡し．**人工知能学会論文誌**, 31(6):764–771, 2016. Computational History: Bridging the Past and Present.

A　時代別可視化結果

本研究で構築した可視化システムによる，時代別の国際関係ネットワークの出力例を示す．各図では，国・都市をノード，共起に基づく関係をエッジとして地図上に配置し，青線が協力関係（Positive），赤線が対立関係（Negative），灰線が中立的関係（Neutral）を表す．

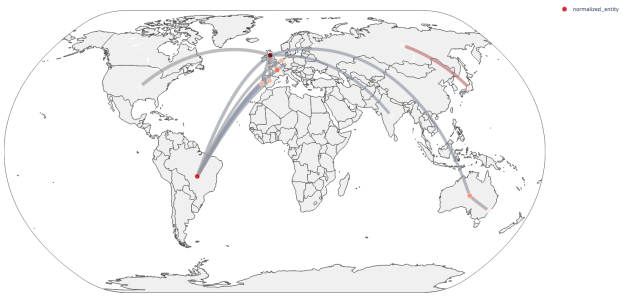

図 2　1500–1799 年の関係ネットワーク．西欧を核に，南北アメリカやアジア・オセアニアなど遠隔地への長距離エッジが目立ち，海外領域をめぐる関係がネットワークの骨格を形成している．

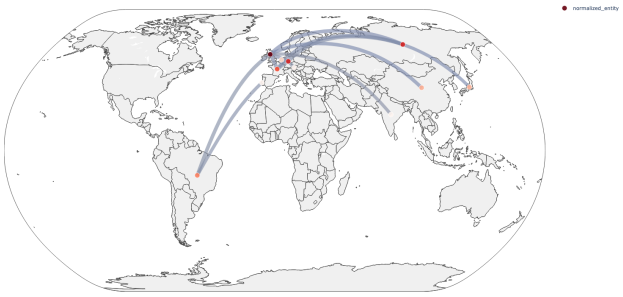

図 3　1800–1899 年の関係ネットワーク．列強間の結び付きが太くなり，欧州を中心とした大陸規模のネットワーク構造へと変化している．

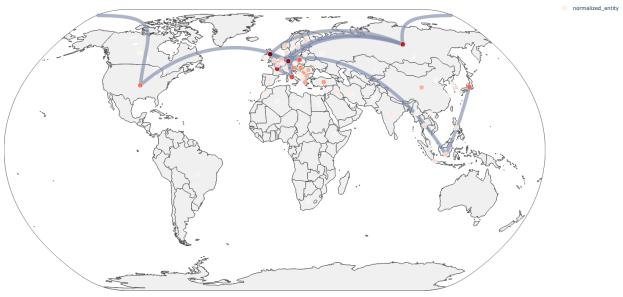

図 4　1900–1945 年の関係ネットワーク．世界大戦期を含むため，欧州周辺の主要国間の結び付きがコアとして凝集し，欧州—北米・欧州—ロシアを結ぶ大西洋・ユーラシアの幹線が際立つ．

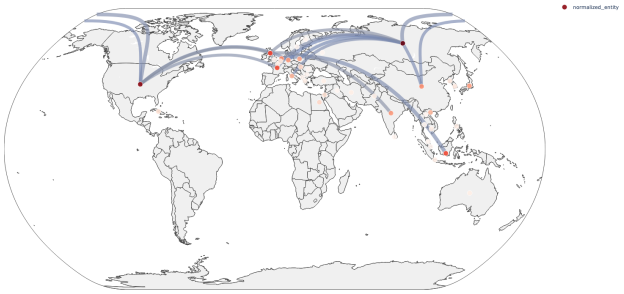

図 5　1946–1969 年の関係ネットワーク．米国とロシア（ソ連）が双ハブとして前面に現れ，二極構造が明確化する．同時に，旧宗主国と独立国間の関係も上位に現れ，脱植民地化に伴う関係の再編が可視化されている．