移動軌跡解析：文章中の人物の地理的な移動を読み取る

山本和太郎 [1]　大友寛之 [2]　大内啓樹 [1,3,5]
東山翔平 [4,1]　寺西裕紀 [3,1]　進藤裕之 [1]　渡辺太郎 [1]
[1] 奈良先端科学技術大学院大学　[2] 株式会社サイバーエージェント
[3] 理化学研究所　[4] 情報通信研究機構　[5] 国立国語研究所
{yamamoto.aitaro.xv6,hiroki.ouchi,shindo,taro}@is.naist.jp
otomo_hiroyuki@cyberagent.co.jp　hiroki.teranishi@riken.jp
shohei.higashiyama@nict.go.jp

概要

本研究では，計算機によって，文章中の人物の地理的な移動を読み取り，その移動軌跡を地図上に再現することをめざす．入力文章中の場所に対し，移動者の訪問状態・訪問順序を予測するタスクを定義するとともに，100 記事からなるアノテーションデータセット ATD-VSO を構築し，ベースラインモデルの学習・性能評価を行った．

1　はじめに

移動軌跡データは，時空間データ解析や地理情報分野において，人間活動と環境条件を分析するための重要データとされている．一般的な移動軌跡は緯度・経度の系列データである．そのため，人間の理解可能な粒度での「場所」の行程として解釈するのは容易でなく，移動者が「東京駅」にいるかどうかという単純な問題さえ判然としない．そこで我々は，文章から移動軌跡を取り出すことを試みる．文章からは，「どの場所にいるか」に加え，その場所で移動者が何をして何を思ったかといった豊かな意味情報も同時に取り出せる可能性を秘める．

このような展望のもと，本研究では，文章中の人物の地理的な移動を読み取る**移動軌跡解析**と，その軌跡を地図上に再現することをめざす．図 1 にその処理過程を示す．移動軌跡解析では，(1) 入力文章から場所を表す表現（場所参照表現）を抽出し，(2) 同じ場所を指す表現をグルーピングする共参照解析を行う．次に，(3) 移動者である文章の書き手が各場所を訪れたか否か等を表す訪問状態を予測し，(4) 訪問場所を時系列順に並べた訪問順序を予測し，移動軌跡を得る．さらに各訪問場所の経緯度を推定するジオコーディングを経ることで，得られた移動軌跡は地図上に可視化可能となる．

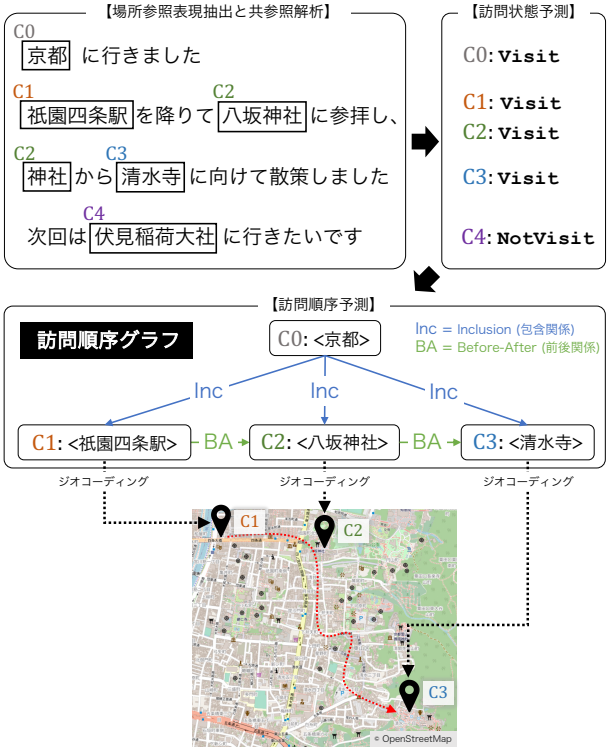

図 1　移動軌跡解析の処理過程と可視化のイメージ

既存研究 [1, 2] では，訪問場所を時系列順に並べた一本の「線」として訪問順序を表現しているものの，「線」では適切に表せない場合も多い（§4.1 で詳述）．そこで本研究では，地理的階層構造も同時に表せる表現形式「**訪問順序グラフ**」を提案する（図 1 中段）．このグラフの情報を 100 記事にアノテーションしたデータセット ATD-VSO[1)] を構築し，訪問状態予測と訪問順序予測についてベースラインモデルの学習・性能評価を行った．

1）**A**rukikata **T**ravelogue **D**ataset with **V**isit **S**tatus and Visiting **O**rder Annotation （https://github.com/naist-nlp/atd-vso）

表 1　エンティティ単位の訪問状態ラベルの一覧

`Visit`	参照先の場所を訪問した.
`VisitPossibly`	参照先の場所を訪問した可能性が示唆されているが不確定.
`Others`	他のいずれの状態にも非該当.

2　前提

「地球の歩き方旅行記データセット」 [3, 4] の一部の国内旅行記に対して，場所参照表現，共参照関係，地理データベース（DB）エントリのリンク情報が付与されたデータセットである ATD-MCL [5] を使用する．ATD-MCL の場所参照表現は，「近鉄奈良駅」のような固有名詞に加え，「お店」「最寄駅」のような一般名詞句も含む．また，同じ場所を指すと解釈できる場所参照表現間には共参照関係が認定され，それらは共参照クラスタをなしている．このようなアノテーション済み旅行記を前提として，本研究では訪問状態と訪問順序に焦点を当てた．

3　訪問状態予測

場所に対する訪問状態を判定する**訪問状態予測**タスクを提案する．たとえば「近鉄奈良駅に到着！」という実体験に基づく記述からは，移動者が「近鉄奈良駅」を訪れたと判断できる．一方，「JR 奈良駅は近鉄奈良駅から少し離れています.」のように事実を説明した記述の場合，移動者がこれらの場所を訪れたことを必ずしも意味しない．このような識別を行うことが本タスクの目的である．

3.1　アノテーションデータ作成

訪問状態の定義　表 1 の 3 種類の訪問状態ラベルを定義し，場所を表す各エンティティ（共参照クラスタ）に対して，適切なラベルを人手で付与した．さらに，各エンティティに含まれる言及に対しても，7 種類の訪問状態ラベル（付録 A に記載）を定義した．エンティティ単位では，文書全体の内容を考慮し，移動者が最終的にその場所を訪れたかを表す訪問状態を決定することが目的である．一方，言及単位では，当該言及の文脈（具体的には文内）から読み取れる内容に基づいて，当該文脈のより詳細な訪問状態を認定することを目的とする．[2)]

作業者間一致率　旅行記 5 記事について，訪問状態ラベルのアノテーションを 2 名の作業者に依頼し，その一致率（F1 値マイクロ平均および Cohen's Kappa κ）を調査した．言及 180 件に対して F1 値 0.80，κ 0.68，エンティティ 124 件に対して F1 値 0.89，κ 0.81 と高い一致率が得られた．

記述統計　1 記事につき作業者 1 名とし，追加の 95 記事にも同様の作業を行い，前述の 5 記事と合わせて，約 3,300 のエンティティを含む 100 記事のアノテーションデータを作成した．詳細は付録の表 6 と表 7 に示す．

3.2　タスク定義

文書の各エンティティ $\boldsymbol{e}_q \in E$ に対して適切な訪問状態ラベル $y \in Y^{\text{vis\_ent}}$ を付与することが目標である．各エンティティ $\boldsymbol{e}_q$ は 1 つ以上の言及 $m_i^{(q)}$ から構成され，$\boldsymbol{e}_q = \{m_1^{(q)}, \ldots, m_{|\boldsymbol{e}_q|}^{(q)}\}$ である．前述のように，各言及にも訪問状態ラベルが定義されているため，まず各言及のラベルを予測し，その予測ラベルをもとにエンティティ単位のラベルを予測するという多段階の解法が可能である．言及単位の予測では，各言及 $m_i^{(q)}$ に対して適切な訪問状態ラベル $y \in Y^{\text{vis\_men}}$ を付与することが目標である．

ベースラインモデル　前述した多段階解法を採用する．まず，各言及 $m_i^{(q)} \in \boldsymbol{e}_q$ に対するラベル確率分布 $P(y|m_i^{(q)})$ を計算し[3)]，これに従って最も可能性の高い訪問状態ラベルを選ぶ．

$$\hat{y}_i^{(q)} = \arg\max_{y \in Y^{\text{vis\_men}}} P(y|m_i^{(q)}) \tag{1}$$

次に，以下のルールに基づいてエンティティ単位のラベルを選ぶ．

1. エンティティ $\boldsymbol{e}_q$ を構成する言及 $m_i^{(q)} \in \boldsymbol{e}_q$ の中で，`Visit` もしくは `PlanToVisit` と予測された言及が 1 つでもあれば，$\boldsymbol{e}_q$ のラベルを `Visit` とする．
2. 上記 1. に該当せず，かつ，`VisitPossibly` と予測された言及が 1 つでもあれば，$\boldsymbol{e}_q$ のラベルを `VisitPossibly` とする．
3. 上記 1. と 2. に該当しない場合は，$\boldsymbol{e}_q$ のラベルを `Others` とする．

4　訪問順序予測

訪問場所の時間的順序関係を判定する**訪問順序予測**タスクを提案する．

2)　「近鉄奈良駅へ.」のように述語が省略された文があることを考慮し，述語にではなく，場所参照表現自体に対して言及レベルの訪問状態ラベルを付与した．

3)　本研究で用いたラベル確率分布相当のモデルは §5 で後述する．

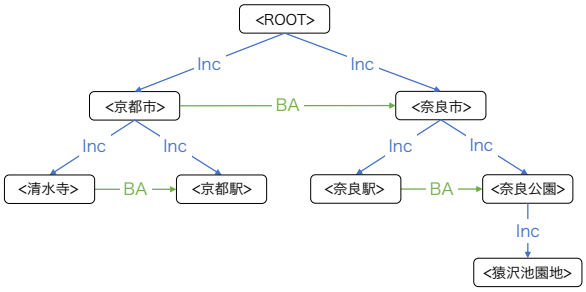

図 2　訪問順序グラフの例．エンティティ間の包含関係を Inc，前後関係を BA で表す．

4.1　目標出力「訪問順序グラフ」

場所間の地理的な**包含関係**と時間的な**前後関係**を考慮した**訪問順序グラフ**を導入する．図 2 の例のように，場所に該当するエンティティがノードであり，エンティティ間にエッジが張られている．包含関係 Inc のエッジは，始点ノードが終点ノードを地理的に包含することを表している．前後関係 BA のエッジは，移動者が始点ノードの場所の次に終点ノードの場所を訪れたことを表している．これらの関係について以下に詳述する．

包含関係　「奈良市に行き，奈良公園を訪れた．」という例文を考える．奈良市の領域が奈良公園の領域を包含しているという地理的知識を踏まえ，「奈良公園を訪れ，その際に必然的に奈良市にも訪れた」と解釈するのが妥当と考える．このように，包含関係にある複数のエンティティをどちらも訪問したことを表現するため，包含関係 Inc を定義する．

前後関係　エンティティ間の包含関係を前提としても，どの階層のエンティティ間に訪問順序を付与すべきかが問題となる．都道府県同士などの同等の粒度で前後関係を認定する素朴な方法は表現力が乏しい[4]．そこで本研究では，「共通の親エンティティを持つエンティティ間」に対してのみ，続けて訪れたことを表す前後関係 BA を付与可能とする．図 2 では，＜奈良市＞を共通の親ノードとする＜奈良駅＞から＜奈良公園＞へ前後関係が付与されている．一方，＜奈良駅＞と＜猿沢池園地＞は親が異なるため，前後関係の付与対象とならない．このように前後関係を表現すると，包含関係のエッジをたどることで，直接的には前後関係が付与されていないエンティティ間についても，どちらを先に訪れたかという情報を得ることが可能になる．

4）　また，「奈良公園 ⊃ 興福寺 ⊃ 興福寺国宝館」のように地域や施設に関して複雑な包含関係が成立するため，あらゆる場所を分類できる階層レベル区分の定義も容易でない．

4.2　アノテーションデータ作成

作業者間一致率　旅行記 5 記事について，関係のアノテーションを 2 名の作業者に依頼し，その一致率（F1 値マイクロ平均）を調査した．前後関係について 0.74[5]，包含関係について 0.94，両関係合わせて 0.85 と概ね高い値であった．

記述統計　1 記事につき作業者 1 名とし，追加の 95 記事にも同様の作業を行い，前述の 5 記事と合わせて，約 3,700 の関係付きエンティティペアを含む 100 記事のアノテーションデータを作成した．詳細は付録の表 8 に示す．

4.3　包含関係予測

タスク定義　文書中で訪問状態が Visit または VisitPossibly のエンティティ集合 $E = \{\boldsymbol{e}_q\}_{q=1}^{|E|}$ の各エンティティ $\boldsymbol{e}_q$ に対して，親エンティティ $\boldsymbol{e}_q^{\mathrm{pa}}$ を予測することが目標である．親エンティティの候補集合を $E_q^{\mathrm{pa\_cand}} = E \setminus \{\boldsymbol{e}_q\} \cup \{\mathrm{ROOT}\}$ と定義する．つまり，エンティティ集合 E から自身 $\boldsymbol{e}_q$ を除き，かつ，他のエンティティを親として持たないことを表す疑似ノード ROOT を加えた集合である．

ベースラインモデル　スコア関数 $\mathrm{score}^{\mathrm{pa}}$ により[6]，候補集合 $E_q^{\mathrm{pa\_cand}}$ から親エンティティである可能性が最も高いエンティティを選出する．

$$\hat{\boldsymbol{e}}_q^{\mathrm{pa}} = \underset{\boldsymbol{e}' \in E_q^{\mathrm{pa\_cand}}}{\arg\max}\ \mathrm{score}^{\mathrm{pa}}(\boldsymbol{e}_q, \boldsymbol{e}') \tag{2}$$

4.4　前後関係予測

タスク定義　文書中の各エンティティ $\boldsymbol{e}_q \in E$ に対して，訪問順序が後続するエンティティ $\boldsymbol{e}_q^{\mathrm{after}}$ を予測する．後続エンティティの候補集合 $E_q^{\mathrm{after\_cand}}$ は，$\boldsymbol{e}_q$ と共通の親を持つエンティティ集合である．

$$E_q^{\mathrm{after\_cand}} = \{\boldsymbol{e}_k | \boldsymbol{e}_q^{\mathrm{pa}} = \boldsymbol{e}_k^{\mathrm{pa}}, \boldsymbol{e}_k \in E\} \cup \{\mathrm{NONE}\}$$

ここで，NONE は後続エンティティを持たないことを表す疑似ノードである．

ベースラインモデル　スコア関数 $\mathrm{score}^{\mathrm{after}}$ により[7]，候補集合 $E_q^{\mathrm{after\_cand}}$ から後続エンティティである可能性が最も高いエンティティを選出する．

$$\hat{\boldsymbol{e}}_q^{\mathrm{after}} = \underset{\boldsymbol{e}' \in E_q^{\mathrm{after}}}{\arg\max}\ \mathrm{score}^{\mathrm{after}}(\boldsymbol{e}_q, \boldsymbol{e}') \tag{3}$$

5）　前後関係の不一致の要因として，一方の作業者が前後関係を認定しているエンティティが他方で包含関係として認定されたことで差異が生じた事例があった．

6）　本研究で用いたスコア関数相当のモデルは §5 で後述する．

7）　本研究で用いたスコア関数相当のモデルは §5 で後述する．

表 2　評価セットにおける訪問状態予測の正解率

手法	言及単位	エンティティ単位
Majority	0.629	0.790
LUKE	0.750	0.838

表 3　評価セットにおける訪問順序予測の F1 値

手法	包含関係	前後関係
Random	0.160	0.206
SimpleOrder	–	0.822
LUKE	0.380	0.744

5　実験設定

タスク設定　訪問状態予測では正解エンティティを所与とした．包含関係予測ではさらに正解訪問状態を所与とした．前後関係予測ではさらに正解包含関係を所与とする設定とし，訪問状態ラベルが `Others` でないエンティティのみ入力とした．

データ分割　100 記事を 7:1:2 で訓練，開発，評価セットに分割した（付録 B に記述統計を示す）．

ベースラインモデル実装　訪問状態予測（式 1）には Hugging Face Transformers[8] の `LukeForEntityClassification` [6] を用いた．包含関係予測（式 2）および前後関係予測（式 3）には `LukeForEntityPairClassification` を用いた．モデルの学習設定などの詳細は付録 C に示す．

6　実験結果

各タスクとも，5 つの異なるシード値で学習させたベースラインモデルの精度の平均を報告する．

訪問状態予測　ベースライン（LUKE）と，訓練データでの出現頻度が最も高いラベルを予測するルール（Majority）による訪問状態予測の正解率を表 2 に示す．言及単位とエンティティ単位の両方において，LUKE が Majority を上回る結果となった．

訪問順序予測：包含関係　ベースライン（LUKE）と，候補エンティティをランダムに選出する方法（Random）による包含関係予測の F1 値（エンティティペア単位）を表 3 に示す．LUKE の精度は Random は超えているものの絶対的には低いと言える．要因として，地域や施設の間の地理的な位置関係が学習されていない点が挙げられる．

訪問順序予測：前後関係　ベースライン（LUKE）と，エンティティを記事中の出現順／ランダムに並べるルール（SimpleOrder ／ Random）による前後関係予測の F1 値（エンティティペア単位，本設定では正解率と一致）を表 3 に示す．LUKE は SimpleOrder より精度が低く，また記事の出現順と逆順序となる前後関係を正しく予測できたケースもなかった．

総括と展望　ベースラインにより，訪問状態予測では正解率 0.84 と実用に向けて期待の持てる精度が得られた．前後関係予測では記事中の出現順に基づくルールに劣る結果となり，エンティティ内の各言及やそれらの周辺文脈を含む大域的な文書内情報を考慮し，出現順と異なる前後関係も捉えられるような改良が必要である．包含関係予測の精度は低いが，エンティティに対応する経緯度や地理 DB エントリを予測するジオコーディングとの併用による改善が考えられる．本研究では，前段タスクの正解を所与としたが，生テキストに対する移動軌跡解析を実現するには，前段タスクの予測結果を基に後段タスクを実施する必要があり，予測誤りの伝搬・累積は避けられない．今後は，個々のタスクの予測精度向上に加え，移動軌跡解析全体を解決可能なシステムの構築に取り組む予定である．

7　関連研究

訪問状態予測　Li ら [7]，松田ら [8] は，ツイートを対象に場所参照表現の抽出とその言及単位での訪問状態の予測を行った．本研究では，同一文書中で同一の場所を指す複数の場所参照表現が出現する状況を想定し，各場所についての最終的な訪問状態としてエンティティ単位での予測も行っている．

訪問順序予測　郡ら [1] は，文章中での地名の出現順を訪問順序と仮定し，ブログからユーザの代表的な行動経路を抽出する手法を提案した．本研究では，出現順と異なる訪問順序も書き手の意図通り正確に抽出することを試みている．Ishino ら [2] は，震災関連ツイートから行動経路の出発地と目的地を抽出する手法を提案した．本研究では，長い文章を想定し，任意の数の訪問場所を含む複雑な行動経路を抽出することに焦点を当てている．

8　おわりに

本研究では，移動軌跡解析に向けて，訪問状態・順序に関するタスク定義，データセット構築，ベースラインモデルの学習・評価を行った．今後，原文を入力として訪問順序までの移動軌跡を予測する移動軌跡解析と，移動軌跡の地図上への接地・可視化までを行うシステムの構築に取り組む予定である．

8）https://huggingface.co/docs/transformers/index

謝辞

本研究は JSPS 科研費 JP22H03648 の助成を受けたものです．本データセットの構築にあたり，「地球の歩き方旅行記データセット」を利用しました．

参考文献

[1] 郡宏志, 服部峻, 手塚太郎, 田島敬史, 田中克己. ブログからのビジターの代表的な行動経路とそのコンテキストの抽出. 情報処理学会研究報告データベースシステム, No. 78 (2006-DBS-140), 2006.

[2] Ishino Aya, Odawara Shuhei, Nanba Hidetsugu, and Takezawa Toshiyuki. Extracting transportation information and traffic problems from tweets during a disaster. The Second International Conference on Advances in Information Mining and Management, 2012.

[3] 株式会社地球の歩き方. 地球の歩き方旅行記データセット, 2022. 国立情報学研究所 情報学研究データリポジトリ.（データセット）. https://doi.org/10.32130/idr.18.1.

[4] 大内啓樹, 進藤裕之, 若宮翔子, 松田裕貴, 井之上直也, 東山翔平, 中村哲, 渡辺太郎. 地球の歩き方旅行記データセット. 言語処理学会第 29 回年次大会発表論文集, 2023.

[5] Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. arXiv:2305.13844, 2023.

[6] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6442–6454, Online, November 2020. Association for Computational Linguistics.

[7] Chenliang Li and Aixin Sun. Fine-grained location extraction from tweets with temporal awareness. In **Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval**, SIGIR'14, p. 43–52, New York, NY, USA, 2014. Association for Computing Machinery.

[8] Koji Matsuda, Mizuki Sango, Naoaki Okazaki, and Kentaro Inui. Monitoring geographical entities with temporal awareness in tweets. In **Computational Linguistics and Intelligent Text Processing: 18th International Conference (CICLing 2017, Revised Selected Papers, Part II 18)**, pp. 379–390, Budapest, Hungary, 2018.

[9] Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. mLUKE: The power of entity representations in multilingual pretrained language models. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 7316–7330, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A　訪問状態ラベルの詳細

言及単位の訪問状態ラベルを表 4 に示す．

表 4　言及単位の訪問状態ラベルの一覧

1	Visit	参照先の場所を訪問した．
2	VisitPossibly	参照先の場所を訪問した可能性が示唆されているが確定しがたい．
3	PlanToVisit	（当該旅行中に）参照先の場所を訪問する予定と言及している．
4	See	参照先の場所を（遠方から）視認したことがわかる．
5	Visit-Past	1〜4 に該当せず，かつ当該旅行以前に，参照先の場所を訪問したことに言及している．
6	Visit-Future	1〜4 に該当せず，かつ当該旅行後に，参照先の場所を訪問する意思があることに言及している．
7	UnkOrNotVisit	参照先の場所を訪問していない可能性が高い，あるいは記述から訪問したことを特定できない．

B　データセット記述統計

本データセットを分割した訓練，開発，評価の各セットの記述統計を表 5 に示す．なお，言及数については，エンティティを構成しているもののみカウントした．言及とエンティティの訪問状態ラベルの分布をそれぞれ表 6（ラベルは表 4 と同一順），表 7（ラベルは表 1 と同一順）に示す．エンティティペア（表 5 の「関係」）のラベルの分布を表 8 に示す[9]．

表 5　本データセットの記述統計

	記事	文	言及	エンティティ	関係
訓練	70	4,254	3,782	2,339	2,598
開発	10	601	505	316	350
評価	20	1,469	1,102	699	777

表 6　言及の訪問状態のラベル別件数

	Visit	Psb	Plan	See	Past	Fut	Unk/Not
訓練	2,390	187	358	212	10	6	619
開発	318	14	48	46	1	4	74
評価	693	55	121	59	10	4	160

C　ベースラインモデル詳細

訪問状態予測　多言語 LUKE [9] 事前学習モデル（https://huggingface.co/studio-ousia/mluke-large-lite）を用いた．対象とする言及を含む文と，文内での言及の位置（文字オフセット）を LukeForEntityClassification の入力として，表 4 の 7 クラスの分類を行った．学習時はエポック数 10，バッチサイズ 16，学習率 5e-6 とした．

表 7　エンティティの訪問状態のラベル別件数

	Visit	Psb	Others
訓練	1,858	84	397
開発	248	4	64
評価	552	23	124

表 8　訪問順序のラベル別件数

	BA	Inc	Unk	Equal	Overlap
訓練	1,037	1,298	125	101	37
開発	138	183	10	10	9
評価	319	375	43	35	5

包含関係予測　日本語 LUKE 事前学習モデル（https://huggingface.co/studio-ousia/luke-japanese-base）を用いた．対象とするエンティティ $\boldsymbol{e}^q$ と各候補エンティティ $\boldsymbol{e}' \in E_q^{\text{pa\_cand}}$ のペアに対して二値分類を行った．その際，$\boldsymbol{e}^q$ 中の代表言及と $\boldsymbol{e}'$ 中の代表言及が出現する各文とそれら 2 文間に出現する文をつないだ文字列，および $\boldsymbol{e}^q$ と $\boldsymbol{e}'$ の各文字オフセットを LukeForEntityPairClassification の入力とした．代表言及は訪問状態ラベルを元に選出した．表 4 の並び順にラベルが Visit の言及を最優先に選択し，さらに，普通名詞よりも固有名詞の言及を優先して選出した．学習時はエポック数 10，バッチサイズ 4，学習率 5e-6 とした．

前後関係予測　包含関係予測と同一の事前学習モデルを用いた．訪問順序グラフ（§4.1）を構築する目標のため，包含関係予測と同様の入力およびハイパーパラメタを使用し，LukeForEntityPairClassification によってエンティティペア単位での前後関係スコアを推定した後，同階層にある全ノードが 1 本の線になるという制約に基づく前後関係系列の推論を行った．具体的には，以下の手順の貪欲探索を用いた．

1. 同一階層にあるエンティティ数 $n = |E_q^{\text{after\_cand}}|$ の集合から得られる ${}_nP_2$ 個のエンティティペア集合 $\mathcal{P}$ のうち，スコア $\text{score}^{\text{after}}$ が最大のペア (e_a, e_b) を選択する．
2. (e_a, e_b) の逆方向のペア (e_b, e_a)，e_b を後件とするペア $(*, e_b)$，e_a を前件とするペア $(e_a, *)$ を元集合から除外した集合 $\mathcal{P}'$ を作る．
3. 上記 2. の手順を繰り返し，全エンティティの前後関係が確定次第終了とする．

9）Unk は前後関係か包含関係かを確定できない関係，Equal は地理的領域がほぼ同一の関係，Overlap は部分的に重なっている関係を表す．