地理的言及に対するエンティティ・リンキングにおける住所階層の利用

三森尊 [1]　乾孝司 [2]
[1] 筑波大学情報学群　[2] 筑波大学大学院 システム情報工学研究群
s2110184@u.tsukuba.ac.jp　inui@cs.tsukuba.ac.jp

概要

地理的位置属性を持つ言及に対するエンティティ・リンキング課題において、地名が持つ住所階層性を活用したエンティティ曖昧性解消手法を提案する。提案手法では、入力文章中に地理的位置属性を持つ言及が複数現れる場合に、それらをトリガにして入力文章を編集することで地名間の住所階層関係をモデルに伝える。評価実験を通して、提案手法は従来手法を上回るエンティティ曖昧性解消の精度を達成し、同名の地名間の混同やカテゴリの異なるエンティティとの混同を効果的に削減できることを確認した。

1　はじめに

エンティティ・リンキングは、テキスト中の言及を Wikipedia[1] や DBPedia[2] などの知識ベース上のエントリ (エンティティ) に関連づける (リンクする) 課題である。エンティティ・リンキングを実施することによって、知識ベースのエントリ情報を考慮した、より高度なテキスト応用処理が可能になる。知識ベースのエントリと関連づけるべき言及は応用ごとに異なるが、文書ジオロケーション [1, 2] や Toponym Resolution [3] のような地理的情報に着目する応用課題では、地名 (「東京都」) やランドマーク (「東京タワー」) といった地理的位置属性を持つ言及が重要となるため、特にこれらの言及を知識ベースのエントリと関連づけることが要求される。

しかしながら、エンティティ・リンキングの既存研究 [4, 5] は主に汎用的なドメインを想定して設計されており、地理的位置属性を持つ言及の特徴を考慮したエンティティ・リンキング手法を提案する。提案手法では、エンティティの扱いに優れた言語モデル LUKE [6] に基づく処理を仮定したうえで、入力文章中の地名が持つ住所階層性に注目して入力事例を事前編集することでエンティティ・リンキングにおけるエンティティ曖昧性解消の性能改善をねらう。

1） https://www.wikipedia.org/
2） https://www.dbpedia.org/

2　構成要素

2.1　地理的言及

本稿の主要な内容に入る前に、本研究の構成要素について説明する。

本研究では、地理的位置属性を持つエンティティへの言及を地理的言及と呼び、地理的言及のみをエンティティ・リンキングの対象として議論を進める。地理的言及の具体的な定義は山本ら [2] の定義を踏襲する。すなわち、関根の拡張固有表現階層 [7] のうち、「組織名」、「地名」、「施設名」および「イベント名」に分類されるエンティティをあらわす言及を地理的言及とする。

2.2　エンティティ・リンキング

エンティティ・リンキングは、テキスト中の言及を知識ベース内のエントリに関連付ける課題であり、言及抽出、候補生成およびエンティティ曖昧性解消の 3 つの副課題に分解して考えることが多い。エンティティ・リンキングは、リンク対象となる知識ベースが Wikipedia 記事である場合は、特に Wikification [8] と呼ばれる。

本研究では地理的言及に対する Wikification を仮定し、さらに副課題のうち、エンティティ曖昧性解消に焦点を当てる。そのため、本研究では言及検出は正しい結果が与えられていると仮定する。また、候補生成には後述する標準的な既存手法を適用することとし、以降では、エンティティ曖昧性解消を中

心に議論する。

2.3 言語モデル LUKE

LUKE [6] は、 RoBERTa [9] をベースとした、エンティティ間の関係抽出や固有表現認識などの課題において高い性能を発揮する言語モデルである。LUKE では、通常の単語とエンティティを区別することができ、エンティティに対してモデルがその意味を深く理解することができる。本研究では、LUKE を言語モデルとして用いた Yamada ら [4] のモデルをエンティティ曖昧性解消モデルに用いる。そのうえで、モデルにわたす文章データを事前編集する手法を提案する。

3 提案手法

後述するベースラインモデルを用いて、地理的言及に対するエンティティ曖昧性解消の誤り分析を実施したところ、以下のような事例において、誤りが頻繁に起こっていることがわかった。

- 地名とは異なるカテゴリでの誤り。例えば、「千葉」という言及に対して、「千葉県」が正解エンティティであるところ、「千葉駅」といった地名以外のカテゴリのエンティティに誤っていた。
- 地名の同名による誤り。例えば、「中央区」という言及に対して、正解エンティティは東京都の「中央区」であるが、札幌市の「中央区」であると誤っていた。

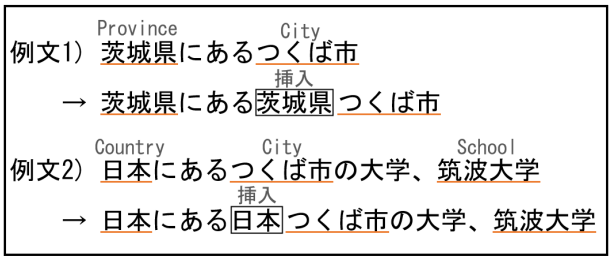
前者の誤りは固有表現認識によって適切なカテゴリ情報が得られれば改善される見込みがある。一方で、後者は地名カテゴリ内での誤りであり、改善には固有表現認識とは別なアプローチが要求される。さらに、上述の例では、入力文章中に「東京都中央区〜」と、人間が見れば正解エンティティが明らかであるような表現がなされていたにもかかわらず、モデルは正しい候補を選択することができていなかった。このことから、ベースラインモデルは地理的言及間の関係を捉える能力が十分ではないことが示唆される。

そこで本研究では、エンティティ曖昧性解消モデルが地理的言及間の関係を捉える能力の向上を目指して、学習時および評価時において、入力文章中に地理的言及が複数現れる場合に、地理的情報を追加するよう入力文章を編集することを考える。より

表 1 階層関係を持つ地名カテゴリ

カテゴリ	言及例
Country	日本, アメリカ合衆国
Province	茨城県, カリフォルニア州
County	東茨城郡, ロサンゼルス郡
City	茨城町, ロサンゼルス

図 1 階層挿入の実行例



具体的には、住所の階層構造に着目して、地名をあらわす文字列を入力文章に挿入する編集をおこなう。以降、本稿ではこれを階層挿入法（upper-level Hierarchical information Insertion, HI）と呼ぶ。図 1 に階層挿入法の適用例を示す。例文１）では、「茨城県」と「つくば市」という２つの地理的言及が含まれているが、この時、「つくば市」は「茨城県」の部分を構成する地名であることが規則的に表されるよう、「つくば市」の直前に地名文字列「茨城県」を挿入する。

具体体な階層挿入法は以下の通りである。入力文章中の任意の地理的言及の対について、以下に示す階層挿入条件を満たした場合に、階層挿入操作を実行する。なお、言及対において、文章中で先行して出現している言及を便宜上、言及 A と称し、それに後続して出現する言及を言及 B とする。

> **【階層挿入条件】** 言及対のどちらの言及も表 1 に示したいずれかのカテゴリに属しており、かつ、言及 B のカテゴリが言及 A よりも住所階層が下位のカテゴリである。
>
> **【階層挿入操作】** 言及 B の直前に言及 A をあらわす文字列を挿入する。

原理的には、「つくば市」と「筑波大学」のような、地名と組織 (の所在地) のような言及対に対しても階層挿入法を適用することは可能である。しかし今回は、議論を単純にするため、表 1 に示すように、地名カテゴリのみに注目して階層挿入を実行することとした。そのため、図 1 の例文２）では、階層挿入条件を満たした「つくば市」に対してのみ挿

入操作がおこなわれている。

LUKE では、文章中の単語（言及）とエンティティの両者を区別した埋め込み表現を利用でき、LUKE を用いてエンティティ・リンキングを実行する場合、リンク処理の対象となる言及に対しては単語埋め込みではなく、エンティティ埋め込みを利用することが通常である。しかし、階層挿入法によって挿入された上位階層にあたる地名言及に対しては単語埋め込みを利用することとした。これによって、挿入操作が実行された下位階層にあたる地名言及に対して、上位階層にあたる地名言及に関する知識を追加的にモデルに渡るようにする。

4 評価実験

4.1 実験設定

提案手法の有効性を検証するために、地理的言及に対するエンティティ曖昧性解消の評価実験を実施した。曖昧性解消モデルには、第 2.3 節で述べたように Yamada ら [4] のモデルを採用するが、日本語データを処理するためにモデル内部で使用する LUKE は日本語版 [3] を用いた。以降では、Yamada ら [4] のモデルを直接使用した場合をベースライン (BL)、文章データに階層挿入法を適用した後に Yamada ら [4] のモデルを使用した場合を階層挿入モデル (HI) とし、両者の性能を比較する。両者の違いは、文章に階層挿入法を適用するかどうかのみであり、エンティティ候補、モデルのアーキテクチャ、およびハイパーパラメータは共通である。学習の詳細設定は付録 A.1 に記載する。

データセットには、日本語 Wikification コーパス [10] を用いた。このコーパスは、拡張固有表現タグ付きコーパス [11] 内の固有表現に対して対応する Wikipedia 記事の ID が付与された Wikification 用のコーパスである。前処理として、対応する Wikipedia 記事がリダイレクト記事である場合、その Wikipedia 記事をリダイレクト先の記事へ置き換えた。このコーパスから以下の条件を全て満たす 5,525 件の固有表現を地理的言及とし、エンティティ曖昧性解消の対象とした。

条件 1. 固有表現のカテゴリが組織名、地名、施設名、イベント名のいずれかである。

条件 2. 固有表現に対応する Wikipedia 記事が候補に含まれる。

候補生成は、陰山ら [12] に倣い、Wikipedia 内で言及がアンカー文字列としてリンクしているエンティティの集合を当該言及のエンティティ候補とした。候補となる記事がリダイレクト記事である場合は、リダイレクト先の記事を候補とした。また、曖昧さ回避ページや一覧ページといった記事は、リンク候補として適切でないと考え候補から除外した。候補生成処理に必要となる Wikipedia 記事データ自体は日本語 Wikification コーパスに含まれていないため、本研究では、2024 年 7 月 1 日時点の日本語 Wikipedia ダンプデータ [4] を使用した。この候補生成手法によって、言及あたり得られる候補エンティティの平均数は 32.9 であった。

階層挿入法を適用するためには地理的言及のカテゴリ情報が必要である。本実験では、学習時は日本語 Wikification コーパスに含まれているカテゴリ情報を参照するが、評価時は LUKE に基づくカテゴリ推定モデルを構築し、カテゴリ情報を推定した。モデルの詳細は付録 A.2 に記載する。

5 分割交差検定を実施し、正解率を評価指標としてエンティティ曖昧性解消の性能を評価した。

4.2 実験結果

実験結果を 表 2 に示す。一行目は、すべての地理的言及に対する結果であり、二行目は、地理的言及のうち、階層挿入が適用された言及（345 件）に絞った場合の結果である。どちらの結果においてもベースラインよりも階層挿入モデルの方が正解率が高く、階層挿入が適用された言及に絞った場合の結果から、その効果が顕著に確認できる。なお、両モデルの結果の間で符号検定を実施し、有意水準 1% で有意差があることを確認している。

表 3 は、すべての地理的言及に対する結果について、文章に含まれる地理的言及数ごとの内訳を見た結果である。この表から、言及数ごとの差はそれ程大きくないことがわかる。文章に含まれる地理的言及の数が 1 つの場合は評価事例に対して階層挿入が発生することはないが、そのような場合でも階層挿入モデルの結果が良いことがわかる。このことから、階層挿入操作がモデル学習時にも良い影響を与えていることが示唆される。

次に、地理的言及のカテゴリごとの正解率を表 4 に示す。階層挿入の適用対象は「地名」カテゴリの

3） https://huggingface.co/studio-ousia/luke-japanese-base

4） https://dumps.wikimedia.org/other/cirrussearch

表 2　実験結果 (正解率)

	BL	HI
全言及	89.8	94.0
HI 言及	82.6	89.6

表 3　文章あたりの地理的言及数ごとの結果

言及数	平均挿入数	BL	HI	言及数
1	0	90.6	94.7	1,609
2	0.125	88.0	94.0	1,326
3 ≤	0.240	90.4	93.6	2590

言及のみであるが、「地名」以外のカテゴリでも性能改善が確認できることは興味深い。

階層挿入が適用された地理的言及に対する改善例を以下に示す。下線が地理的言及を示し、このうち、注目している言及を太字で示している。

- 【正解：宮田町 (福岡県) ／ BL 出力：宮田町 (愛知県)】
  Province City
  福岡県 **宮田町** の同社内での会見で「暖かい
  Domestic_Region
  九州　で花を咲かせたい。」

この事例に階層挿入法を適用すると「宮田町」の直前に「福岡県」が挿入され、HI モデルへの入力事例は「福岡県福岡県宮田町の〜」と変化し、この編集による追加情報によって、モデルは宮田町が福岡県に属するということを正しく推定できたと考えられる。階層挿入後は表層上は「福岡県」という語が連続しているため冗長に見えるが、第 3 節で述べたように階層挿入前の入力事例における「福岡県」は曖昧性解消の対象となるため LUKE モデルはこれを未知エンティティとして扱うが、挿入された「福岡県」は単語として扱われる。これにより入力事例の地理的な文脈が強化され正解に至ったと考えられる。

また、階層挿入が入力事例中で発生しなかった場合においても改善が見られた。

- 【正解：地中海／ BL 出力：地中海食】
  Sea Island
  地中海 の シチリア島 に育ち、十二歳で料理の道を志した。

階層挿入が発生しなかった改善例の傾向として、BL ではカテゴリの異なるエンティティ間で混同していたような事例に対して多く改善が見られた。このような事例の改善は、モデル学習時の階層挿入の適用により、曖昧性解消モデルがある程度の地理的知識を獲得したことが曖昧性解消の結果に反映されたものであると考えられる。

表 4　カテゴリごとの正解率

カテゴリ	BL	HI	言及数
組織名	90.3	93.9	3,466
地名	89.4	94.9	1,383
イベント名	86.9	91.9	397
施設名	90.0	95.0	279

階層挿入法では、言及情報は参照せず、カテゴリ間の階層関係のみに基づいて挿入が実行されるため、現実にはエンティティが地理的包含関係にない場合でも適用される。そのため、以下のような誤りが発生することも確認された。

- 【正解：ローマ／ HI 出力：1987 年世界陸上選手権大会】
  Country City
  ロシア は 4 日、**ローマ** で合同会議を開き・・・

この例では、BL で正しいエンティティを出力することができたが、階層挿入法で「ローマ」の直前に「ロシア」が挿入されることで、 HI では誤ったエンティティが出力された。

以下の事例は、 BL でも HI でも正解エンティティを出力することができなかった例である。この事例では「ニュージランド」と「南島」は現実では階層関係をもつが、「南島」が “Island” カテゴリとして扱われるため階層挿入法が適用されなかった。

- 【正解：南島 (ニュージーランド) ／ BL, HI 出力：ユージヌィ島】
  Country Island
  ・・・ニュージランド の **南島** にあるアベル・タスマン国立公園。

5　おわりに

本研究では、地理的言及に特化したエンティティ・リンキングのための、地名が持つ階層性に注目した入力事例の編集手法を提案し、評価実験の結果からこの手法が地理的言及に対するエンティティ曖昧性解消に対して有効性があることを示した。特に地名へのエンティティ曖昧性解消に対して、同名の地名や異なるカテゴリのエンティティとの混同に階層挿入法が有効に機能することがわかった。

本稿で述べた階層挿入条件を満たす事例は言及全体のわずか 6%（345/5,525）であった。そのため、今後は適用範囲の拡大について検討する必要がある。また、実験結果の誤り例で示したように、階層挿入法の改善として、階層挿入条件への地理的包含関係の組み込みについて検討する必要がある。

参考文献

[1] Han Bo, Cook Paul, and Baldwin Timothy. “Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pp. 1045–1062, 2012.

[2] Yuya Yamamoto and Takashi Inui. Utilizing Geographic Entity Information for PLM-based Document Geolocation Models. In *The 38th Pacific Asia Conference on Language, Information and Computation*, 2024.

[3] Jochen L. Leidner. “Toponym Resolution in Text: “Which Sheffield is it?”. in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 602, 2004.

[4] Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. “Global Entity Disambiguation with BERT.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3264- -3271, 2022.

[5] Satoshi Sekine, Kouta Nakayama, Maya Ando, Yu Usami, Masako Nomoto, Koji Matsuda, and Asuka Sumida.“SHINRA2020-ML: Categorizing 30-Language Wikipedia into Fine-Grained NE Based on Resource by Collaborative Contribution Scheme.” *In Proceedings of the 3rd Conference on the Automated Knowledge Base Construction*, 2021.

[6] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. “LUKE: Deep contextualized entity representations with entity-aware self-attention.”
In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6442–6454, 2020.

[7] Satoshi Sekine and Yoshio Eriguchi. “Extended Named Entity Hierarchy.” In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’ 02)*, pp. 1818–1821, 2002.

[8] Rada Mihalcea and Andras Csomai. “Wikify! Linking documents to encyclopedic knowledge.” In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 233–242, 2007.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. “Roberta: A robustly optimized BERT pretraining approach.” arXiv preprint arXiv:1907.11692, 2019.

[10] Davaajav Jargalsaikhan, 岡崎直観, 松田耕史, 乾健太郎. 日本語 Wikification コーパスの構築に向けて. 言語処理学会第 *22* 回年次大会発表論文集, pp. 793–796, 2016.

[11] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会自然言語処理研究会 *(2008-NL-188)*, pp. 113–120, 2008.

[12] 陰山 宗一, 乾 孝司. 文書ジオロケーション課題における地理的特定性指標の有効性評価. 自然言語処理, Vol.31, No.4, pp.1665–1690, 2024

[13] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. “8-BIT OPTIMIZERS VIA BLOCK-WISE QUANTIZATION”. arXiv preprint arXiv:2110.02861, 2022.

A　付録

A.1　曖昧性解消モデルの詳細

第 4 節で述べた本研究のベースとなるエンティティ曖昧性解消モデルの詳細を示す。このモデルは、日本語版事前学習済み LUKE モデルとエンティティ予測ヘッドの 2 つで構成される。

エンティティ曖昧性解消モデルは入力として文章、文章中の曖昧性解消対象となる言及および言及に対応した関連付けエンティティ候補を受け取る。事前学習済み LUKE モデルは、この入力から各地理的言及に対応した埋め込み表現を獲得する。その埋め込み表現をエンティティ予測ヘッドで線形変換とソフトマックス関数を適用後、 Wikipedia 内の全エンティティに対するスコアに変換する。このとき、言及の候補エンティティではないエンティティのスコアを非常に小さな値に設定することで、予測エンティティとして選択されないようにする。最後に、候補エンティティの中でスコアが最も高いエンティティをその言及に対応するエンティティとして選択して出力する。最適化手法として 8-bit AdamW [13] を使用し、損失関数として交差エントロピー誤差を用いた。第 4 節の評価実験で用いたベースラインモデルと階層挿入モデルで用いるハイパーパラメータを表 5 に示す。

表 5　エンティティ曖昧性解消モデルの学習設定

ネットワーク	パラメータ	値
共通	バッチサイズ	1
	エポック数	2
LUKE	最大入力トークン長	512
	語彙数	32,772
	エンティティ語彙数	570,505
	隠れ層サイズ	768
	ドロップアウト率	0.1
	学習率	2e-5
	重み減衰	0.01
	AdamW β_1	0.9
	AdamW β_2	0.999
	AdamW ϵ	1e-6
予測ヘッド	入力次元	768
	出力次元	570,505
	学習率	2e-5

A.2　カテゴリ推定モデルの詳細

第 4 節で構築した、階層挿入法で使用するカテゴリ推定モデルの詳細を示す。このモデルのアーキテクチャは、日本語版事前学習済み LUKE モデルをベースとしている。事前学習済み LUKE モデルはエンティティ曖昧性解消モデルと同様に、文章中の言及に対する埋め込み表現を獲得する。この埋め込み表現をカテゴリ予測ヘッドで、線形変換とソフトマックス関数を適用後、データセット内で使用される 53 種類のカテゴリに対するスコアに変換し、最もスコアの高いカテゴリを予測カテゴリとして出力する。エンティティ曖昧性解消モデルと同様に、最適化手法として 8-bit AdamW を使用し、損失関数として交差エントロピー誤差を使用した。このモデルの学習で用いるハイパーパラメータを表 6 に示す。表 6 に示されていないパラメータは、エンティティ曖昧性解消モデルと同一である。

表 6　カテゴリ推定モデルの学習設定

ネットワーク	パラメータ	値
共通	バッチサイズ	4
	エポック数	3
LUKE	学習率	1e-5
予測ヘッド	出力次元	53
	学習率	1e-5

カテゴリ推定モデルの性能評価として、5 分割交差検定を実施し、正解率を評価指標としてカテゴリ推定モデルの性能を評価した。結果を表 7 に示す。

表 7　カテゴリ推定モデルの精度

地理的言及数	5,525
正解数	5,223
正解率 (%)	94.5