日本語旅行記ジオパージングデータセット ATD-MCL

東山翔平 [1,2]　大内啓樹 [2,3]　寺西裕紀 [3,2]　大友寛之 [4]
井手佑翼 [2]　山本和太郎 [2]　進藤裕之 [2]　渡辺太郎 [2]
[1] 情報通信研究機構　[2] 奈良先端科学技術大学院大学　[3] 理化学研究所　[4] CyberAgent, Inc.
shohei.higashiyama@nict.go.jp　hiroki.ouchi@is.naist.jp
hiroki.teranishi@riken.jp　otomo_hiroyuki@cyberagent.co.jp
{ide.yusuke.ja6, yamamoto.aitaro.xv6, shindo, taro}@is.naist.jp

概要

ジオパージングは，文章中に含まれる実世界の場所に関わる情報を解析する基盤技術である．本研究では，文書レベルのジオパージングシステムの開発・評価に向けた第一歩として，日本語旅行記データセットの構築と現在のシステムの評価を行った．本データセットが，日本語ジオパージング研究の推進・高度化に貢献することを期待している．

1　はじめに

場所を指し示す言語表現（**場所参照表現**）は，人間の活動・移動や実世界で生じた事象の描写と共に文章に記述されることが多い．言語表現と実世界の場所を紐づける**ジオパージング** [1, 2] は，テキストを，地理的な観点に関わる様々な目的で活用するのに欠かせない技術である．文章として書かれた事物・事象に関する情報の地図上への可視化，地理情報や地図を介したテキスト検索が可能になるほか，観光支援，災害対応，感染症サーベイランスなどの応用領域での活用も期待されている [3]．

ジオパージングは，場所参照表現の抽出（ジオタギング）と，場所参照表現が指す場所を特定する曖昧性解消（ジオコーディング）の二つのサブタスクから構成される．ジオパージングでは，個々の場所参照表現を独立に解析することも可能であるものの，他の場所参照表現の情報を考慮した文書レベルの解析をすることで，高度な曖昧性解消が可能となる．というのも，同一文書中で複数の場所参照表現が共起する際，それらの情報は相互に役立つ場合があるためである．たとえば，図 1 ⟨3⟩ の「興福寺」は，奈良県への旅行に関する文脈を考慮することで，他地域の興福寺ではなく奈良県の興福寺を指していると判定できる．

近鉄奈良駅 FAC-NAME ⟨1⟩ に到着。そこ DEICTIC ⟨1⟩ から
奈良公園 FAC-NAME ⟨2⟩ までは歩いてすぐです。
お寺 FAC-NOM ⟨GENERIC⟩ が好きなので最初に興福寺 FAC-NAME ⟨3⟩
に行きました。境内 FAC-NOM ⟨3⟩ で鹿と遭遇し、
奈良 LOC-NAME ⟨4⟩ に来たことを実感しました。

⟨1⟩ https://www.openstreetmap.org/relation/11532920
⟨2⟩ https://www.openstreetmap.org/way/456314269
⟨3⟩ https://www.openstreetmap.org/way/1134439456
⟨4⟩ https://www.openstreetmap.org/relation/3227707

図 1　アノテーション済み文書の例．場所参照表現とそのエンティティ種別（例「FAC-NAME」），共参照クラスタ ID（例「⟨1⟩」），共参照クラスタに対応する OpenStreetMap エントリ URL の情報が付与されている．

本稿では，我々が構築・公開したジオパージングデータセット ATD-MCL[1] について報告する．本データセットは，日本語の旅行記に対し，3 種類の地理的な情報，つまり (1) 場所参照表現，(2) それらの共参照関係，(3) 共参照クラスタから地理データベース（DB）エントリへのリンク情報を人手でアノテーションしたものである（図 1）．我々の知る限り，第三者が入手・再現可能な日本語ジオパージングデータセットは他にはない．

本データセットは文書レベルジオパージングに適した二つの特徴を持つ．1 点目は，1 文書中に場所参照表現が豊富に含まれることである．原文書が有するこの性質をモデル学習・評価時に活用可能にするため，様々な言語表現で表された様々なカテゴリの場所を指す場所参照表現を網羅的にアノテーションする方針を採用した．なお，この性質は SNS 投稿のような短いテキストを使用した既存データセット [4, 5] とは対照的である．2 点目は，1 文書中の場所参照表現の間に，共参照関係や，参照先の場所の

[1] **A**rukikata **T**ravelogue **D**ataset with Geographic Entity **M**ention, **C**oreference, and **L**ink Annotation (https://github.com/naist-nlp/atd-mcl)

地理的近さ・階層的関係といった地理的関連性があることである．旅行記は，旅行者の移動が時系列的に記される文書の典型と捉えられるため，ニュース記事を用いた既存データセット [6, 7, 8, 2] よりも顕著にこの特徴を持つと考えている．

本データセットを用いて現在のエンティティ解析システムの精度を評価したところ，場所参照表現抽出と共参照解析については高い精度が得られた一方，曖昧性解消においては改善の余地が大きいことが示された．

2 アノテーションデータの構築

設計方針 原文書に出現する様々な場所参照表現（以降，**メンション**と呼ぶ）に対して網羅性の高いアノテーションを実現するため，二つの方針を設けた．1 点目に，場所・施設に関連する広範なカテゴリの種別ラベル（§2.1 で後述）を定義し，固有名に限らずそれら種別に該当する表現をアノテーション対象とした．2 点目に，施設名に対しても高い割合で DB エントリのリンク情報を付与できるよう，施設エントリを豊富に含む地理 DB として OpenStreetMap（OSM）[2] を採用した．

データ準備 原データとして，「地球の歩き方旅行記データセット」[9, 10] を用いた．同データ中で国内旅行記に分類されている文書のうち，文書長が長すぎないもの（500–3000 文字に相当する文書）に限定して 200 記事をランダムに抽出し，更に日本語自然言語処理ライブラリ GiNZA [3] [11] による固有表現抽出を適用し，作業用データとした．

アノテーション手順・ツール アノテーション委託先企業の日本語母語アノテータ（各ステップ 5〜7 名）により，(1) メンション，(2) 共参照，(3) リンクの三つのアノテーションステップを順に実施した．アノテーションツールとして，(1) と (2) では brat [4] [12]，(3) では Microsoft Excel を使用した．

2.1 メンションアノテーション

本ステップでは，自動付与結果を修正し，入力文書中のメンションのスパンと**種別ラベル**を特定する作業を行った．種別ラベルとして，表 1 に示す 8 種類のラベルを定義した．LOC，FAC，LINE，TRANS は，固有名の表現であるか，その他の表現であるかに応じてサブ種別 NAME か NOM が付与される．LOC_ORG と FAC_ORG はメトニミー的に組織を指す地名と施設名に対するラベル，LOC_OR_FAC は地域・施設両方を指し得る非固有名の表現に対するラベル，DEICTIC は場所を指す指示表現に対するラベルである．

表 1 場所参照表現の種別ラベルと例

概要	種別-サブ種別	例
地域・地形	LOC-NAME LOC-NOM	奈良; 生駒山 町; 島
施設	FAC-NAME FAC-NOM	大神神社 駅; 公園
線状の地物	LINE-NAME LINE-NOM	近鉄奈良線 国道; 川
乗り物	TRANS-NAME TRANS-NOM	特急ひのとり バス; フェリー
組織を指す地域名	LOC_ORG	
組織を指す施設名	FAC_ORG	
地域・施設不定	LOC_OR_FAC	観光地; ところ
指示表現	DEICTIC	そこ

2.2 共参照アノテーション

本ステップでは，メンションへの**特定性ラベル** GENERIC およびメンション対への**関係ラベル** COREF を付与する作業を行った．GENERIC は，総称的な表現（例：図 1 の「お寺」）に付与するラベルであり，実世界の場所を指すが他メンションと共参照関係を持たないシングルトンと区別するために用いた．COREF は，実世界の同一の場所を指すメンション対（例：図 1 ⟨1⟩ の「近鉄奈良駅」と「そこ」）に対して付与するラベルである．ラベル付与作業の後，COREF の 2 項関係を介して連鎖的に繋がりを持つメンション全体は一つの共参照クラスタ，いずれのラベルも付与されていないメンションはシングルトン（例：図 1 ⟨2⟩ および ⟨4⟩）として扱った．

2.3 リンクアノテーション

本ステップでは，OSM 検索およびウェブ検索結果を参照しながら，共参照クラスタに対して，その参照先の場所に相当する OSM エントリ URL（例：図 1 の ⟨1⟩–⟨4⟩）を割り当てる作業を行った．参照先の場所に直接的に相当せず，その場所を包含するようなエントリのみ見つかった場合には，同エントリの URL と共に PART_OF ラベルを付与した．適切なエントリが見つからない場合には，URL の代わりに NOT_FOUND を割り当てた．

[2] https://wiki.openstreetmap.org/
[3] https://github.com/megagonlabs/ginza
[4] https://github.com/nlplab/brat

表 2　ATD-MCL の記述統計

	#Doc	#Sent	#Word	#Mention	#Cluster
Set-A	100	5,949	85,741	6,052	3,131
Set-B	100	6,324	87,074	6,119	3,208
全体	200	12,273	172,815	12,171	6,339

2.4　データセットの記述統計

200 記事にメンション・共参照アノテーションを実施し，そのうち 100 記事にリンクアノテーションを追加で実施した[5]．2 種類の情報のみ付与された 100 記事を Set-A データ，3 種類の情報が付与された残り 100 記事を Set-B データとする．構築したアノテーションデータの記述統計を表 2 に示す．単語数の計測には SudachiPy [13] モード B を用いた．

3　システム性能評価

本データセットを用いて，メンション抽出（MR），共参照解析（CR），エンティティ曖昧性解消（ED）の 3 タスクの実験を行った．本実験の目的は，既製公開モデル，fine-tuning モデルを含む現在のエンティティ解析システムによって達成可能な解析精度の水準を明らかにすることである．

データ分割　Set-B データ 100 記事を 10:10:80 で訓練，開発，テストセットに分割した．MR および CR では更に Set-A データ 100 記事を訓練セットに加え，110:10:80 記事のデータを使用した．ED では Set-B データの 10:10:80 記事を使用した．

3.1　メンション抽出

タスク定義　MR タスクを，入力文書からメンションのスパンと種別ラベルを特定する問題とする．評価指標には，正解メンションのスパン・種別ラベルとの完全一致の F1 値を用いる．

システム　ATD-MCL 訓練セットで fine-tuning した 2 システム（spaCy-MR，mLUKE-MR）と，既製公開モデルの 2 システム（KWJA，GiNZA）を評価した．spaCy-MR は，日本語 ELECTRA [14] 事前学習モデル[6]をベースに自然言語処理ライブラリ spaCy の遷移型モデル[7] を fine-tuning したシステムである．mLUKE-MR は，多言語 LUKE [15] 事前学習モデル[8]をベースに，我々が実装したスパン抽出型モデルを fine-tuning したシステムである．KWJA（version 2.1.1） [16] では base モデル，GiNZA（version 5.1.2）では ja_ginza_electra モデルを用いた．

表 3　MR システムの精度（NAME，NOM は，それらサブ種別ラベル全体に対する精度のマイクロ平均）

システム	種別	適合率	再現率	F1 値
KWJA	Overall	.279	.352	.311
	NAME	.279	.695	.398
GiNZA	Overall	.574	.277	.374
	NAME	.574	.548	.560
spaCy-MR	Overall	.752	.732	.742
	NAME	.733	.719	.726
	NOM	.798	.763	.780
mLUKE-MR	Overall	**.813**	**.817**	**.815**
	NAME	.828	.813	.821
	NOM	.832	.826	.829

実験結果　各 MR システムのテスト精度を表 3 に示す．GiNZA と KWJA は，NAME メンションに対し，ある程度のカバレッジ（再現率 0.55–0.70）を示した．一方，これらは固有表現に特化して学習されたモデルであるため，一般名詞句・指示表現など NAME でないメンションに対しては抽出に失敗している[9]．spaCy-MR と mLUKE-MR は，fine-tuning のおかげで F1 値 0.74–0.82 と比較的高い精度を示した．両システムとも，NOM に比べて NAME に対する精度が低いのは，NAME の方が表層の多様性が大きく，認識が難しいためと考えられる．

3.2　共参照解析

タスク定義　CR タスクを，正解メンションを所与とし，共参照関係にあるメンション同士を同一クラスタにまとめ上げる問題とする．評価指標には MUC [17]，B^3 [18]，$CEAF_e$ [19]（各 F1 値）と，それら F1 値のマクロ平均である CoNLL スコア [20] を用いる．

システム　ATD-MCL 訓練セットで fine-tuning したシステム（mLUKE-CR），既製公開モデル（KWJA），二つのルールベースシステム（Rule-CR-1 および 2）を評価した．mLUKE-CR は，多言語 LUKE 事前学習モデル[10]をベースに，我々が実装した先行詞予測型の end-to-end CR モデル [21] を fine-tuning したシステムである．KWJA では base モデルを使用し，出

[5] 他の 2 種類のアノテーション作業の約 3 倍の作業時間を要することから，100 記事の作業量にとどまった．
[6] https://huggingface.co/megagonlabs/transformers-ud-japanese-electra-base-discriminator
[7] https://spacy.io/api/architectures#parser
[8] https://huggingface.co/studio-ousia/mluke-large-lite
[9] NOM に対する各精度は 0 である（表 3 には不記載）．
[10] https://huggingface.co/studio-ousia/mluke-large

表 4 CR システムの精度（クラスタサイズ別）

システム	サイズ	MUC	B^3	$CEAF_e$	Avg.
Rule-CR-1	≥ 1	0	.755	.639	.465
	≥ 2	0	0	0	0
Rule-CR-2	≥ 1	.622	.840	.790	.750
	≥ 2	.622	.613	.629	.621
KWJA	≥ 1	.694	.839	.793	.775
	≥ 2	.694	.661	.658	.671
mLUKE-CR	≥ 1	**.753**	**.875**	**.839**	**.822**
	≥ 2	**.753**	**.733**	**.737**	**.741**

力クラスタの和集合を正解メンション集合に一致させる後処理を適用した．Rule-CR-1 は，すべてのメンションをシングルトンとみなす．Rule-CR-2 は，表層文字列が同一のメンションを一つのクラスタとし，その他のメンションはシングルトンとみなす．

実験結果　各 CR システムのテスト精度を表 4 に示す．シングルトンが多いデータの分布の表れとして，「何もしない」Rule-CR-1 でも，全クラスタ（サイズ ≥ 1）に対して 0.6 以上の B^3 および $CEAF_e$ スコアとなった．Rule-CR-2 では各指標で 0.61–0.84 の精度となり，単純な表層情報が共参照関係を捉えるのに有効であることを示している．KWJA および mLUKE-CR の精度は，これら 2 システムが表層の異なるメンション間の共参照関係もある程度捉えられていることを示している．MR の結果と同様，fine-tuning を行っている mLUKE-CR が最も高い精度を示した．

3.3　エンティティ曖昧性解消

タスク定義　ED タスクを，正解エンティティ（＝共参照クラスタ）を所与とし，各エンティティに対して全 DB エントリの中から適切なエントリを選択する問題とする[11]．評価指標には Recall@k，つまり，システムが出力した k 個のエントリが正解エントリを含んでいれば正解とする尺度を用いる．

システム　教師なし ED システム（BERT-ED）およびルールベースシステム（Rule-ED）を評価した．両システムとも，入力エンティティに対し，表層文字列が最長の NAME メンションをエンティティの「名前」として扱い，それに基づいて DB エントリを割り当てる[12]．BERT-ED は，日本語 BERT [22] 事前学習モデル[13]を用いたシステムであり，名前文字列を基にしたエンティティベクトル，name 属性値文字列を基にした DB エントリベクトル[14]を計算し，コサイン類似度が上位 k 件の DB エントリを出力する．Rule-ED では，エンティティの名前と name 属性値が完全一致する DB エントリを抽出し，エントリの全属性値を連結した文字列を辞書式順序でソートして上位 k 件を出力する．

[11] 厳密には，同一視可能な DB エントリをグループにまとめ，全 1.8M エントリグループの中から適切なグループを選択を行うタスクとした．詳細は付録 A に示す．

[12] NAME メンションを含まない入力エンティティに対してはエントリの推定を行わない．

表 5 ED システムの精度（“R” は Recall を表す）

システム	R@1	R@5	R@10	R@100
Rule-ED	.221	.323	.345	.362
BERT-ED	.219	.366	.399	.482

実験結果　各 ED システムのテスト精度を表 5 に示す．ベクトルに基づくソフトなマッチングにより，$k > 1$ で BERT-ED の方が高い精度を示し，k が大きい時に精度差が顕著である．このように BERT のベクトル表現が有用である結果は示されたものの，k が小さい時の精度は改善の余地が大きい．

3.4　議論

MR および CR では fine-tuning システムが高い精度を達成した一方，ED で用いた教師なしシステムは低い精度にとどまった．改善方法の一つは，教師ありシステムの学習，つまり fine-tuning を実施することであるが，未知事例に対する予測精度が主な課題になると考える．もう一つの課題は，Wikipedia のような DB と異なり，OSM 含む地理 DB では自然言語の説明文が利用できないことである．しかし，OSM エントリは意味的な属性情報を有し，またエントリ間の距離や包含関係などの地理的な関係も取得可能である．エンティティや DB エントリのベクトル表現の学習にこのような地理的な関係を取り入れつつ，複数のメンションに関する文書レベルの情報を考慮するような発展が考えられる．

4　おわりに

本稿では，文書レベルジオパージングに利用可能な公開アノテーションデータセット ATD-MCL の構築と評価について報告した．本データセットが，日本語ジオパージング技術の研究開発の推進・発展に貢献することを期待している．

[13] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

[14] 入力文字列中の各トークンについての BERT 最終層の出力ベクトルの平均を用いた．

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．本研究の過程でご助言いただいた松田裕貴氏，若宮翔子氏，井之上直也氏，山田育矢氏に感謝申し上げます．

参考文献

[1] Jochen L Leidner. An evaluation dataset for the toponym resolution task. Computers, Environment and Urban Systems, Vol. 30, No. 4, pp. 400–417, 2006.

[2] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. A pragmatic guide to geoparsing evaluation: Toponyms, named entity recognition and pragmatics. Language resources and evaluation, Vol. 54, pp. 683–712, 2020.

[3] Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. Location reference recognition from texts: A survey and comparison. ACM Comput. Surv., Vol. 56, No. 5, 2023.

[4] Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. Geographical entity annotated corpus of japanese microblogs. Journal of Information Processing, Vol. 25, pp. 121–130, 2017.

[5] Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. Geocorpora: building a corpus to test and train microblog geoparsers. International Journal of Geographical Information Science, Vol. 32, No. 1, pp. 1–29, 2018.

[6] Michael D. Lieberman, Hanan Samet, and Jagan Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In ICDE, pp. 201–212. IEEE, 2010.

[7] Ehsan Kamalloo and Davood Rafiei. A coherent unsupervised model for toponym resolution. In WWW, p. 1287–1296, 2018.

[8] Milan Gritta, Mohammad Taher Pilehvar, and Nigel Collier. Which Melbourne? Augmenting geocoding with maps. In ACL, pp. 1285–1296, July 2018.

[9] Arukikata. Co., Ltd. Arukikata travelogue dataset, 2022. Informatics Research Data Repository, National Institute of Informatics. https://doi.org/10.32130/idr.18.1.

[10] Hiroki Ouchi, Hiroyuki Shindo, Shoko Wakamiya, Yuki Matsuda, Naoya Inoue, Shohei Higashiyama, Satoshi Nakamura, and Taro Watanabe. Arukikata travelogue dataset. arXiv:2305.11444, 2023.

[11] 松田寛, 大村舞, 浅原正幸. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会第 25 回年次大会発表論文集, 2019.

[12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In EACL, pp. 102–107, 2012.

[13] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a Japanese tokenizer for business. In LREC, 2018.

[14] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR, April 2020.

[15] Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. mLUKE: The power of entity representations in multilingual pretrained language models. In ACL, pp. 7316–7330, 2022.

[16] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A unified japanese analyzer based on foundation models. In ACL, Toronto, Canada, 2023.

[17] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In MUC-6), 1995.

[18] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, Vol. 1, pp. 563–566, 1998.

[19] Xiaoqiang Luo. On coreference resolution performance metrics. In HLT-EMNLP, pp. 25–32, October 2005.

[20] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In EMNLP-CoNLL Shared Task, pp. 1–40, 2012.

[21] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In EMNLP, pp. 188–197, September 2017.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, pp. 4171–4186, 2019.

[23] Satoshi Sekine and Hitoshi Isahara. IREX: IR & IE evaluation project in Japanese. In LREC, 2000.

[24] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In LREC, 2002.

[25] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In EMNLP-IJCNLP, pp. 5803–5808, 2019.

A　データベース前処理

OSM の原データベース[15]は非常に多数のエントリを含み，更に実質的に同一の場所を指すエントリも多く含む[16]．そこで，同一視できるエントリを一つのグループにまとめ上げる前処理を行い，ED の実験では個々のエントリではなくエントリのグループを単位として予測する設定を採用した．

前処理は次のように行った．まず，日本国内の場所を表す 2.8B エントリから，name 属性値を持つ 2.8M エントリを抽出した．次に，name 属性値に加えて住所や一部の属性値の情報を連結した文字列[17]を key とし，key が同一のエントリを同一グループとした．その結果，1.8M グループとなった．

B　システム詳細

§3 で示した fine-tuning 済みシステムの精度は，いずれも 1 回の実行結果の精度である．

メンション抽出　spaCy-MR では，GiNZA の ja_ginza_electra モデルとほぼ同様のモデル設定とした[18]．mLUKE-MR は，LUKE Transformer モデルを用いてスパン（メンション候補）の分散表現を計算し，全結合層を用いてスパンをいずれかの種別ラベルまたは非メンションに分類した．入力文（512 トークンに達するまで周囲のトークンを連結）中の可能なすべてのスパンについて，それぞれスパンの先頭・末尾トークンの分散表現とエンティティの分散表現を連結してスパンの分散表現とした．KWJA の IREX [23] タグ，GiNZA の拡張固有表現 [24] タグの出力結果に対して，ATD-MCL のタグへ変換するルールを適用した[19]．

共参照解析　mLUKE-CR では，Joshi ら [25] のモデル構造に基づき，mLUKE-MR と同様の方法でメンションの分散表現を計算した[20]．KWJA へは，出力クラスタの和集合を正解メンション集合に一致させる後処理[21]を適用した．

モデルハイパーパラメタ　mLUKE-MR/CR のモデル学習で用いたハイパーパラメタは，両モデル共通で linear learning rate decay，warmup ratio=0.06，dropout rate=0.1，weight decay rate=0.01，gradient clipping なし，Adam $\beta_1 = 0.9$，$\beta_2 = 0.98$，$\epsilon = 1e-6$ とし，MR/CR モデル別で学習率 1e-5/5e-5，バッチサイズ 8/4，学習エポック数 10/20 とした（学習率と学習エポック数のみパラメタ探索を行った）．spaCy-MR では，事前学習モデル（§3.1）に加えて後述する設定ファイルの値を使用した．BERT-ED では事前学習モデル（§3.3）からの追加のハイパーパラメタはない．

C　アノテータ間一致率

メンション，共参照，リンクのアノテーションにおいて，各 10 記事，10 記事，5 記事について 2 人目のアノテータによるアノテーション作業を実施し，1 人目のアノテータとの間の一致率を算出した．後述のように，3 種類の情報いずれについてもアノテーションの一貫性がある程度高いことを示す一致率を得た．

メンションアノテーションの一致率として，一方の作業結果を正解とした場合の F1 値を算出したところ，NAME メンションでは 0.835，NOM メンションでは 0.867，全メンションでは 0.832 となった．

共参照アノテーションの一致率として，一方の作業結果を正解とした場合の CoNLL スコア（F1 値）を算出した．全クラスタに対して 0.858，サイズ 2 以上のクラスタに対して 0.792 となった．

リンクアノテーションの一致率として，2 名の作業結果（エントリ URL または NOT_FOUND の割り当て）についての κ 係数と，一方を正解とした場合の F1 値を算出した．(a) そのままの URL で比較する設定では κ 係数 0.707，F1 値 0.722 となり，(b) 実質的に同一の場所を指す URL を同一視する設定（同一視可能かは第一著者が判断）では κ 係数 0.793，F1 値 0.804 となった．

[15] http://download.geofabrik.de/asia/ にて公開されていた japan-230601.osm.bz2 を使用した．

[16] 我々のタスクで同一視可能な例として，name 属性値「東京」を持つエントリ 72 件の中に，同一鉄道駅の異なるホームや異なる列車停止位置を指す複数のエントリがある．

[17] 例：“name=スターバックス|branch=None|prefecture=奈良県|city=奈良市|quarter=樽井町|road=猿沢遊歩道|amenity=cafe”

[18] https://github.com/megagonlabs/ginza/blob/develop/config/ja_ginza_electra.cfg（“transformer” および “ner” 以外の pipeline は無効にした．）

[19] KWJA の出力 LOCATION メンションに対しては，同一スパンでラベルを LOC_NAME，FAC_NAME，LINE_NAME のいずれかに変更した計 3 インスタンスに複製する再現率重視の処理を行った．

[20] ただし，メンションの unary スコア，coarse-to-fine 推論，メタデータに基づく離散素性は使用しなかった．

[21] 正解メンションとスパンに重なりのある予測メンションは，正解メンションと同一視した．その上で，出力クラスタからどの正解メンションとも一致しない予測メンションを削除し，どの予測メンションとも一致しない正解メンションをシングルトンとして出力に加えた．