日英対訳ジオパージングデータセット ATD-Para

東山 翔平 [1,2]　大内 啓樹 [2,3]　藤田 篤 [1]　内山 将夫 [1]
[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
{shohei.higashiyama,atushi.fujita,mutiyama}@nict.go.jp hiroki.ouchi@is.naist.jp

概要

テキストに含まれる地理情報の解析技術は，観光・防災支援などの応用に有用である．本研究では，日本語旅行記の英訳とアノテーションにより，日・英 2 言語の地理的解析・生成タスクのデータセットを構築した．基本的な 3 タスクにおいて，言語（日・英）および地域（国内・海外）の観点でシステムの性能評価・分析を行った．

1　はじめに

地理情報は，実世界のモノ／対象物に関する位置情報（経緯度，住所など）と非位置情報（属性情報）が対になった情報である [1]．たとえば，「あるオブジェクトが，“奈良県生駒市” にあり，その通称名が “往馬大社” である」という情報は地理情報である．

テキストは，人間の視点から実世界のモノの情報が描写された，暗黙的な地理情報データとも捉えられる．テキストの顕著な特徴として，(i) 内在する位置情報（**場所参照表現**）の非明示性，(ii) 非位置情報の種類・内容の豊富さ[1]，(iii) ユーザ生成テキストを含めたウェブデータの量の多さが挙げられる．テキストデータは，(ii) や (iii) の特徴から，様々な地理的応用 [2] への活用が期待できる．一方，(i) の点から，場所参照表現を明示的な位置情報へ変換する**ジオコーディング**（曖昧性解消）処理が必要であり，同処理を高精度に行うことは課題の一つである．

地理的なテキストの解析・生成について，多言語情報アクセスの観点では，原文や機械生成文の記述言語や話題の対象となる地域にかかわらず，各言語のユーザが同様に高品質な解析結果や生成テキストを享受できることが理想と言える．しかし，様々な要因から，言語や地域によって言語処理モデルの性能に差があることが想定され，このことは，情報アクセスにおける格差を生む．たとえば，非日本語話者が，日本語でしか書かれていないような日本の地域の情報を知りたい場合，チャットボット・機械翻訳システム等の対象言語の処理（翻訳）性能が低いほど，日本語話者や，相対的に性能が高い言語の話者に比べて不利となる．

こうした背景から，本研究では，旅行記を用いた日・英 2 言語の地理的解析・生成タスク評価データセット ATD-Para[2] を構築した．日本国内・海外の旅行についての日本語原記事に対し，その英訳を作成し，場所参照表現と位置情報（OpenStreetMap [3] エントリ URL）を付与した．これにより，同等の内容の 2 言語のテキストに対する (1) 場所参照表現抽出と (2) ジオコーディング，場所参照表現を豊富に含むテキストの (3) 機械翻訳という，地理的な言語処理に関する基本的な 3 タスクでのモデル学習・言語横断評価を可能とした．

これら 3 タスクについて，記述言語（日・英）と地域（日本国内・海外）の違いに対するモデルの性能差を調査することを目的として，各種 Transformer [4] モデルを用いたシステムの性能評価実験を行った．

なお，機械翻訳や人手翻訳を利用し，同内容の多言語テキストの場所参照表現抽出結果を比較した既存研究 [5, 6] は存在する．本研究では，複数言語でのジオコーディングや地理情報を考慮した機械翻訳の評価を行った点などが新規である．

2　データセットの構築

地球の歩き方旅行記データセット（ATD） [7, 8] は，旅行記投稿サイトに投稿された日本語旅行記原文データである．ATD の日本国内旅行記を用いたアノテーションデータセットに ATD-MCL[3] [9] がある．付与された情報は，場所参照表現（メンション），それらの共参照関係，メンションが指す位置

1　たとえば，様々な対象物について，Wikipedia 記事本文に含まれる非構造情報に比べて，Infobox や Wikidata 上の構造化情報が大幅に少ないことからも実感できる．

2　本データセットは，国立情報学研究所 情報学研究データリポジトリ（https://www.nii.ac.jp/dsc/idr/）より，「地球の歩き方旅行記翻訳データセット」として公開する．

3　https://github.com/naist-nlp/atd-mcl

表 1 ATD 原文および二次データの記事数.

	原文 (Raw)	地理 (Geo)	翻訳 (Tra)
国内旅行記 (Dom)	4,500	200	58
海外旅行記 (Ovs)	9,500	78	32

に相当する地理データベース OpenStreetMap（OSM）エントリのリンク情報の 3 種類である.

本研究では，(a) ATD の海外旅行記に，ATD-MCL と同様の情報を付与した海外旅行記地理情報データ，(b) ATD 国内・海外旅行記の原文を英語に翻訳した翻訳文データを作成し，(c) 両者を統合した対訳ジオパージングデータセット ATD-Para を構築した.

ATD および二次データの関係と記事数を表 1 に示す．Raw 全体の 1.4 万記事が ATD に相当し，Dom-Geo の 200 記事が ATD-MCL，Ovs-Geo の 78 記事がデータ (a), Tra 全体の 90 記事がデータ (b), Tra 全体の 90 記事とそれらに対応する {Dom, Ovs}-{Raw, Geo} を合わせたものが提案データ (c) である.

2.1 海外旅行記地理情報データ Ovs-Geo

Ovs-Geo は，ATD の海外旅行記に 3 種類の情報を付与したデータであり，独立したデータとしても公開している[4]．Dom-Geo と同様のアノテーション基準・方法 [9] に則り，メンション・共参照関係付与についてアノテータ 2 名，リンク付与について 5 名により作業を行った（アノテータはいずれも委託先企業の日本語母語話者）．データサイズは，78 記事，4,313 文，5,116 メンションとなった.

本データの全 78 記事中 5 記事に対しては，アノテータ 2 名により独立にメンション・共参照関係を付与し，同 5 記事中 3 記事に対しては，アノテータ 2 名により独立にリンク情報を付与した．これら 5 または 3 記事に付与されたメンション，共参照関係，リンク情報についての 2 名のアノテータ間一致率は，それぞれ F1 値 0.90，LEA [10] スコア 0.86，Cohen's κ 0.69 と，概ね高い一致を示した.

2.2 翻訳文データ {Dom, Ovs}-Tra

{Dom, Ovs}-Tra は，ATD の国内・海外旅行記を翻訳したデータである．翻訳作業は，委託先翻訳会社にて ISO 17100:2015 [11] に従ったワークフローに基づき，日本語母語の日英翻訳者 2 名，英語母語のバイリンガルチェッカー 3 名により行った．翻訳を行った記事は，表 1 に示す計 90 記事である.

翻訳作業仕様は，簡潔にまとめると次のようにな

4 https://github.com/naist-nlp/atd-mcl-overseas

表 2 実験データの記述統計（日本語 / 英語）.

	#Sec.	#Sentence	#Men_Tag	#Men_QA
Dom-Train	421	1,120 / 1,100	542 / 540	415
Dom-Dev	199	440 / 434	212 / 211	165
Dom-Test	330	1,023 / 971	430 / 430	322
Ovs-Train	238	869 / 919	532 / 534	464
Ovs-Dev	131	409 / 414	185 / 185	169
Ovs-Test	190	611 / 689	332 / 331	285

る．(1) セクション[5]を単位とし，英語として自然な表現・情報構造の文章となるよう，各記事を翻訳者 1 名が翻訳する．(2) 日本語原文は，場所参照表現が所定のタグで囲まれたテキストとして与えられ，訳文中で対応する表現は同一のタグで囲む．(3) 場所参照表現を含む固有表現について，正式・一般的な英語表記がある場合にはそれを使用する．同一記事中で同一固有表現を 2 回以上訳出する場面では，適宜，代名詞等を使用したり省略しても良い.

3 システム性能評価

本データセットを用いた実験について報告する．§1 で述べたように，本実験の目的は，地理情報が関わる 3 タスクにおいて，最近のモデルの性能を評価し，記述言語（日・英）と地域（日本国内・海外）の違いによる傾向を分析することである.

国内および海外旅行記のデータサイズを同程度とするため，国内旅行記 58 記事中 32 記事と，海外旅行記 32 記事のそれぞれを，15，6，11 記事に分割し，順に訓練，開発，テストセットとして用いた[6]．各言語の各セットの記述統計を表 2 に示す.

3.1 場所参照表現抽出

本タスクでは，モデルの場所参照表現抽出精度を評価する．入力文から固有名の地名（LOC_NAME）・施設名（FAC_NAME）のメンションを特定する問題とし，表 2 に示すメンション数（#Men_Tag）となった．正解メンションとのスパンの完全一致の F1 値（ラベルの一致は考慮しない）を評価指標とし，予測メンションの正確さを評価する.

システム Masked LM（MLM）に基づく既製の固有表現抽出器として，spaCy [12] en_core_web_trf 3.0.0，GiNZA [13] ja_ginza_bert_large β1 を用いた．また，スパン抽出ツール LUKE-NER[7] [9] を用いて，

5 文章と画像の対からなる旅行記投稿者による記事内の区切りを指す．翻訳時に画像は参照しない.

6 国内旅行記の訓練，開発，テストセットは，ATD-MCL [9] の実験で使用されたそれらのそれぞれサブセットである.

7 https://github.com/naist-nlp/luke-ner

表 3 場所参照表現抽出精度（テストセット）．下線は各ブロック内での最高スコア．Fine-tuned モデル（†）では 3 回の実行の平均スコアを示す．

	Prec.	Rec.	F1	Prec.	Rec.	F1
System	Ja-Dom			Ja-Ovs		
GiNZA	0.633	0.737	0.681	0.719	0.756	0.737
mLUKE†	0.878	0.788	0.828	0.841	0.840	0.841
Llama^{8B}	0.331	0.353	0.342	0.373	0.190	0.251
Swallow^{8B}	0.398	0.835	0.539	0.486	0.786	0.601
Llama^{70B}	0.619	0.381	0.472	0.780	0.512	0.618
Swallow^{70B}	0.692	0.867	0.770	0.714	0.813	0.761
System	En-Dom			En-Ovs		
spaCy	0.753	0.730	0.741	0.773	0.804	0.788
mLUKE†	0.876	0.849	0.862	0.808	0.881	0.843
Llama^{8B}	0.717	0.723	0.720	0.743	0.689	0.715
Swallow^{8B}	0.527	0.809	0.639	0.640	0.731	0.683
Llama^{70B}	0.866	0.723	0.788	0.813	0.698	0.751
Swallow^{70B}	0.764	0.835	0.798	0.761	0.749	0.755

多言語 LUKE（mLUKE） [14] 事前学習モデル[8]の fine-tuning を行い[9]，評価した．さらに，指示学習済み Causal LM（CLM）として，英語中心の Llama-3.1 [15]（8B[10]，70B[11]）と，Llama-3.1 に継続事前学習を施した日英対応モデルの Llama-3.1-Swallow [16]（8B[12]，70B[13]）を評価した．各 CLM では，Language Model Evaluation Harness [17] を利用し，付録 A のプロンプトで 10-shot（10 文）文脈内学習を実施した．

結果と議論　各システムの精度を表 3 に示す．Fine-tuning 済み mLUKE が 4 データとも最高精度（F1 値 0.85 前後）であり，言語・地域による明確な精度差は見られなかった．CLM では，10-shot 学習ながら，Swallow-70B が日・英両データで高い精度（F1 値 0.75 以上）を示した．同モデルも地域の違いによる明確な精度差は見られなかった．目立った点として，Llama は日本語・国内（Ja-Dom）で低精度だが，英語では国内・海外（En-Dom/Ovs）とも良好な精度を示した．これは，Llama の日本語表記の（特に国内の）地名認識能力が乏しい一方で，英語表記では海外と同等以上に日本国内の地名・施設名の特徴を捉える能力を有することを示唆している．

8　studio-ousia/mluke-large-lite
9　言語ごとに Dom-Train と Ovs-Train を統合したデータを学習データとし，バッチサイズ 32，エポック数 8，その他の設定は文献 [9] と同様にし，各言語のモデルを学習させた．
10　meta-llama/Llama-3.1-8B-Instruct
11　meta-llama/Llama-3.1-70B-Instruct
12　tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1
13　tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1

3.2　ジオコーディング

本タスクでは，モデルのメンション曖昧性解消精度を評価する．CLM も評価可能な問題設定として，対象メンションが出現する入力文 1 文に対し，選択肢から適切な OSM エントリを選ぶ四択問題とした[14]．三つの不正解選択肢には，対象メンションと名前が類似するエントリまたはランダムに選択されたエントリを採用した[15]．Dom，Ovs データについて，正解エントリ情報を有するメンションのうち，それぞれ日本国内，国外の場所に当たる事例に限定した結果，表 2 に示す設問数（#Men_QA）となった．

システム　ルールベース法として，Exact Match 法（EM）と Fuzzy Match 法（FM）を評価した．EM，FM はそれぞれ，メンション文字列とエントリ名文字列の完全一致，表層類似度[16]により回答候補を得て，候補中からランダムに回答する[17]．MLM ベースの方法として，RoBERTa [18] 日本語モデル（Large[18]; “RoBERTa-J” と呼ぶ），XLM-R [19]（Large[19]）の事前学習モデルを用いたシステムを評価した．具体的には，入力文と各選択肢テキストをベクトル表現に変換後，両者の内積を計算し，内積最大の選択肢を回答とした．CLM として，§3.1 と同一の 4 モデルを評価した．各 CLM では，付録 B のプロンプトを用いて 4-shot（4 問）文脈内学習を行い，各選択肢番号（“1”〜“4”）の対数尤度により回答を決定した．

結果と議論　各システムのテスト精度（正解率）を表 4 に示す．FM の正解率は各データで 5 割前後である．事前学習 MLM 手法では，RoBERTa-J が日本語データでのみ FM 以上の精度を示した．CLM の 4 モデルは全般的に高い正解率を示した．期待と異なり，Llama は日本語データでも同サイズの Swallow と同等以上の精度を示し，各言語のデータでモデル間の明確な精度差は見られなかった．これは，日本語地名知識が不十分でもある程度解決可能な問題形式であったことが一因と考えられる．

14　本実験は，可能な全エントリからの候補生成とリランキングの 2 ステップのうち後者に相当する．全エントリから正解を予測する設定での実験は，今後の課題とする．
15　詳細は次の通り．正解メンション文字列を用いたクエリで，OSM 対応ジオコーダ Nominatim Search API に問い合わせてレスポンスを取得．次に，レスポンス上位から最大 3 件を選択し，3 件未満の場合はランダムなエントリを加えた．
16　https://pypi.org/project/fuzzywuzzy/, ratio() を使用．
17　両ルールベース法では，回答候補数の逆数（回答候補数 0 の場合は値 0）の平均により正解率の期待値を算出した．なお，文字列先頭の “The”/“the”を除外後に計算した．
18　nlp-waseda/roberta-large-japanese
19　FacebookAI/xlm-roberta-large

表 4 ジオコーディング正解率（テストセット）

System	Ja-Dom	Ja-Ovs	En-Dom	En-Ovs
Exact Match	0.305	0.237	0.282	0.401
Fuzzy Match	0.513	0.467	0.484	0.505
RoBERTa-J^{337M}	0.534	0.530	0.193	0.172
XLM-R^{560M}	0.339	0.232	0.286	0.298
Llama^{8B}	0.742	0.839	0.823	0.814
Swallow^{8B}	0.755	0.846	0.789	0.797
Llama^{70B}	0.792	0.884	0.845	0.832
Swallow^{70B}	0.808	0.888	0.832	0.835

また，4 モデルとも，国内データでは日本語，海外データでは英語の方が精度が低い点は，四択問題データの作成方法に起因するデータの特徴が影響したとみられる[20]．

3.3 機械翻訳

本タスクでは，モデルのセクション単位での日英翻訳精度を評価する．評価指標には，BLEU[21] [21]，COMET（`wmt22-comet-da`）[22]，Term Success Ratio（TSR）[23] を用いた．TSR は，原文中の各場所参照表現に対する参照訳中の訳語を，システム出力が含んでいるかを Fuzzy match で評価する指標である[22]．

システム 従来 NMT モデルとして，NLLB-200（600M，3.3B）[24] を評価した．CLM として，§3.1 と同一の 4 モデルを評価した．各 CLM では，(a) 単純なプロンプトと，出現メンションの用語対訳を (b) 正解 OSM エントリ日・英名称または (c) 参照訳中の訳語により指定したプロンプト（付録 C）を用いて，4-shot（4 セクション）文脈内学習を行った．

結果と議論 各システムのテスト精度を表 5 に示す．従来 NMT に比べ，CLM が高い精度を示した．CLM の (a) の結果は，同サイズでは Llama よりも Swallow が高精度であり，継続事前学習での日・英対訳文使用の効果とみられる．地域差については，モデル問わず海外データで高精度であったが，データの特徴が影響したとみられ[23]，訳語の妥当性評価についてはより頑健な方法も必要と考えられる．(b)，(c) の結果からは，文脈内学習時の用語対訳指定の有効性が確認できる．(b) でも TSR を中心に一定の効果が見られるが．正解訳語を指定した (c) での顕著な向上と比べると限定的であり，ノイズを含む知識ベース情報を取捨選択しながら活用することなどが必要と言える．

表 5 日英翻訳精度（テストセット）．(a) からの 1（5）ポイント以上のスコア向上／低下：↑／↓（↑↑／↓↓）．

System	Dom			Ovs		
	BLEU	COM	TSR	BLEU	COM	TSR
NLLB^{600M}	6.8	59.2	4.2	9.8	67.9	36.6
NLLB$^{3.3B}$	12.5	67.8	9.3	16.1	74.0	52.6
(a) CLM with Basic Prompt						
Llama^{8B}	22.3	77.7	35.0	26.6	82.8	59.8
Swallow^{8B}	25.8	80.2	51.4	30.8	84.4	68.9
Llama^{70B}	26.7	80.5	49.1	31.7	84.4	71.0
Swallow^{70B}	28.7	81.4	59.2	34.1	85.3	71.6
(b) CLM with Bilingual Entry Name Prompt						
Llama^{8B}	23.4↑	78.9↑	44.4↑↑	28.6↑	83.3	67.7↑↑
Swallow^{8B}	25.9	80.1	55.0↑	33.0↑	84.7	71.0↑
Llama^{70B}	28.7↑	80.8	55.8↑↑	30.0↓	84.2	71.3
Swallow^{70B}	29.6	80.3↓	58.3	34.1	84.7	72.8↑
(c) CLM with Oracle Term Prompt						
Llama^{8B}	29.3↑↑	81.5↑	95.0↑↑	30.1↑	84.5↑	94.6↑↑
Swallow^{8B}	30.2↑↑	82.2↑	95.6↑↑	33.4↑	85.0↑	93.4↑↑
Llama^{70B}	33.4↑↑	83.2↑	95.0↑↑	35.5↑	85.6↑	96.1↑↑
Swallow^{70B}	34.3↑↑	82.4↑	96.0↑↑	36.6↑	85.8	94.9↑↑

4 おわりに

本研究では，地理的解析・生成タスクのためのデータセット ATD-Para を構築し，2 言語（日・英）および 2 地域（国内・海外）のデータを用いたモデル性能評価を行った．言語間の比較では，英語中心 Llama モデルはタスクによっては日本語精度が低くなる一方，2 言語の単言語・対訳データで事前学習された Swallow モデルは安定的に高い精度を達成するなどの結果であった．地域間の比較では，モデルごとに明確な精度差がないか，差が生じたケースはデータの特徴に起因するとみられるものであった．

今後も，より実用的・統合的なタスク設定での評価を含め，多言語のテキスト地理情報処理技術の開発・評価を進める予定である．

20 同傾向であった開発データでのモデルの誤り事例を分析したところ，国内地名は日本語データ，海外地名は英語データの方が，不正解選択肢に占める類似名エントリ数が多い事例が散見され，それぞれ他方の言語のデータよりも難しい問題を多く含む側面があったとみられる．これは，問題作成時，Nominatim API へ問い合わせたメンション文字列の言語（＝データの言語）が，取得された不正解選択肢用エントリの内容・件数に影響したためである．

21 sacreBLEU [20]，signiture="`nrefs:1|case:mixed|eff:no|tok:intl|smooth:exp|version:2.4.3`".

22 https://pypi.org/project/fuzzywuzzy/, `partial_ratio()`

23 同傾向が見られた開発データで，モデル出力中の場所参照表現を分析したところ，国内外データとも (i) 固有名中の一般名詞語句の訳し方の不一致（"Mt. Akadake" vs. "Akadake"），(ii) 読み・綴りの誤り，の 2 種のエラーが共通して多く見られたが，(i) を生じるような一般名詞語句を持つ事例が国内データにより多く存在したことの影響があると考えられる．

謝辞

本研究の一部について，JSPS 科研費 23K24904 の助成を受けました．データセットの構築・提供にあたり，株式会社地球の歩き方の曽我将良氏，株式会社 Gakken の上原康仁氏，国立情報学研究所の大須賀智子氏，大山敬三氏から多大なご協力をいただいたことを深謝します．

参考文献

[1] 浅見泰司, 矢野桂司, 貞広幸雄, 湯田ミノリ. 地理情報科学 GIS スタンダード. 古今書院, 2015.

[2] 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, 2022. arXiv:2207.01683, 2022.

[3] OpenStreetMap contributors. OpenStreetMap, 2015. https://www.openstreetmap.org.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS**, p. 6000–6010, 2017.

[5] Judith Gelernter and Wei Zhang. Cross-lingual geo-parsing for non-structured data. In **GIS**, p. 64–71, 2013.

[6] Xu Chen, Judith Gelernter, Han Zhang, and Jin Liu. Multi-lingual geoparsing based on machine translation. **Future Gener. Comput. Syst.**, Vol. 96, No. C, p. 667–677, July 2019.

[7] 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. 国立情報学研究所 情報学研究データリポジトリ. https://doi.org/10.32130/idr.18.1, 2022.

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

[9] 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. In **Findings of the ACL: EACL 2024**, pp. 513–532, March 2024.

[10] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? A proposal for a link-based entity aware metric. In **ACL**, pp. 632–642, August 2016.

[11] ISO/TC37. ISO 17100:2015 translation services—Requirements for translation services, 2015. https://www.iso.org/standard/59149.html.

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020. https://www.doi.org/10.5281/zenodo.1212303.

[13] 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.

[14] Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. mLUKE: The power of entity representations in multilingual pretrained language models. In **ACL**, 2022.

[15] Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

[16] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **COLM**, pp. 1–25, October 2024.

[17] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2024. https://zenodo.org/records/12608602.

[18] 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. In **ICLR**, pp. 1–15, 2020.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **ACL**, pp. 8440–8451, July 2020.

[20] Matt Post. A call for clarity in reporting BLEU scores. In **WMT**, pp. 186–191, October 2018.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **ACL**, pp. 311–318, July 2002.

[22] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In **WMT**, pp. 578–585, December 2022.

[23] Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. Findings of the WMT 2023 shared task on machine translation with terminologies. In **WMT**, pp. 663–671, December 2023.

[24] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. arXiv:2207.04672, 2022.

A　場所参照表現抽出の実験詳細

既製システムの実験設定　spaCy では，GPE，LOC，FAC ラベルの出力事例を採用した．GiNZA では文献 [9] §D.2 のラベル変換ルールを適用した．

CLM の実験設定　Swallow モデルでは図 1 のような日本語プロンプトを使用した．ただし，指示文直後に挿入される 10-shot 事例（入出力対 10 文）は記載を省略した．Llama モデルでは，日本語プロンプトと概ね同内容の英語プロンプトを使用した．

入力文から、地名・地形名（ラベル：LOCATION）および施設名（ラベル：FACILITY）に該当する固有表現をすべて抽出し、”固有表現:ラベル;固有表現:ラベル;・・・”の形式で、出現順に列挙してください。該当するものがなければ”NONE”のみ出力してください。\n 入力：(省略)\n 出力：

図 1　場所参照表現抽出用 CLM 日本語プロンプト

B　ジオコーディングの実験詳細

多肢選択問題データ　入力文は，図 2 のプロンプトと同様，メンション部分を「⟨⟨」と「⟩⟩」（半角山括弧各 2 個）で囲んで表示した．正解・不正解選択肢は，Nominatim Lookup/Search API により OSM エントリ属性情報を取得し，図 2 のプロンプト中の選択肢のように name，display_name（“full_name” と表記），category，type，addresstype 属性を用いた文字列表現を生成した．この際，レスポンス言語をクエリ中で指定することで，name および display_name 属性の日本語と英語両方の値を取得し，日本語データでは日本語名，英語データでは英語名を用いた[24]．

MLM の実験設定　MLM 手法の詳細を記す．入力文と各選択肢テキストを，事前学習 MLM の最終層出力ベクトルを用いてそれぞれベクトル表現に変換した．各テキストのベクトル化では，[CLS] トークンのベクトルを用いる方法（CLS）とメンションスパン内トークンのベクトルの平均を用いる方法（AVG）を試行し，また選択肢テキストについては，前述のテキスト表現全体を入力する方法（Entire）と full_name 属性値のみ入力する方法（FullName）を試行し，開発セットで性能が良かった CLS かつ FullName を採用した．回答は，§3.2 に記載の通り，入力文ベクトルとの内積の値により決定した．

CLM の実験設定　CLM の実験について，Swallow モデルでは図 2 のような日本語プロンプトを使用した（指示文直後に挿入される 4-shot 事例は記載略）．

24　ただし，指定した言語の name 属性値がない場合は，異なる言語（対象 POI の現地語など）の値が返却される．

Llama モデルでは，前述の日本語プロンプトと同内容で，指示文を英語，選択肢中の name，full_name の値を英語名称にしたプロンプトを使用した．

入力テキスト中の⟨⟨場所参照表現⟩⟩が指している場所を、与えられた選択肢（1,2,3,4）の中から選んでください。\n
入力：以前から⟨⟨出雲⟩⟩に行きたいと思っていましたが、中国山地を列車で越えて見たいとの思いもあり、岡山経由で山陰に入ることにしました。
選択肢：
1: name=”出雲市”. full_name=”出雲市, 島根県, 日本”. category=boundary. type=administrative. addresstype=city.
2: name=”出雲”. full_name=”出雲, 穂波嘉穂線, 大字九郎丸, 飯塚市, 嘉穂郡, 福岡県, 日本”. category=highway. type=traffic_signals. addresstype=highway.
3: name=”出雲”. full_name=”出雲, 亀岡園部線, 千歳町千歳, 亀岡市, 京都府, 621-0001, 日本”. category=highway. type=bus_stop. addresstype=highway.
4: name=”出雲”. full_name=”出雲, 国道 165 号, 黒崎, 桜井市, 奈良県, 633-0017, 日本”. category=highway. type=traffic_signals. addresstype=highway.
回答：

図 2　ジオコーディング用 CLM 日本語プロンプト．入力文は ATD [7] 旅行記の実例（開発データ事例）を引用．

C　機械翻訳の実験詳細

CLM の実験では，各モデル共通で図 3 のような英語プロンプトを使用した（指示文直後の 4-shot 事例は記載略）．

Please translate the following Japanese text into English.\n
Japanese: 1 日目 以前から出雲に行きたいと思っていましたが、中国山地を列車で越えて見たいとの思いもあり、岡山経由で山陰に入ることにしました。(以降省略)
English:

Please translate the following Japanese text into English. Repalce each ⟨⟨location expression⟩⟩ in the text with the appropriate English term, referring to the provided mappings (format: ⟨⟨original source expression⟩⟩ - ⟨⟨ source term || target term⟩⟩) when relavant. If a mapping is not relevant, translate the location expression normally without using the mappings. Ensure that the ⟨⟨ ⟩⟩ brackets remain around each translated term.\n
Term mappings: ⟨⟨出雲⟩⟩ - ⟨⟨出雲市 || Izumo⟩⟩;⟨⟨岡山⟩⟩ = ⟨⟨岡山市 || Okayama⟩⟩;(以降省略)
Japanese: 1 日目 以前から⟨⟨出雲⟩⟩に行きたいと思っていましたが、⟨⟨中国山地⟩⟩を列車で越えて見たいとの思いもあり、⟨⟨岡山⟩⟩経由で山陰に入ることにしました。(以降省略)
English:

Please translate the following Japanese text into English. Repalce each ⟨⟨location expression⟩⟩ in the text with the appropriate English term as specified in the provided mappings (format: ⟨⟨source term ⟩⟩ = ⟨⟨target term⟩⟩), ensuring that the ⟨⟨ ⟩⟩ brackets remain around each translated term.\n
Terms: ⟨⟨出雲⟩⟩ = ⟨⟨Izumo⟩⟩;⟨⟨中国山地⟩⟩ = ⟨⟨Chugoku Mountains⟩⟩;⟨⟨岡山⟩⟩ = ⟨⟨Okayama⟩⟩;(以降省略)
Japanese: (省略)
English:

図 3　翻訳用 CLM プロンプト．上段から順に §3.3 のプロンプト (a)，(b)，(c) に該当．