LLM から抽出した日本文化知識のデータベース構築と活用

大橋 巧 [1]　彌冨 仁 [1]
[1] 法政大学大学院 理工学研究科 応用情報工学専攻
takumi.ohashi.4g@stu.hosei.ac.jp　iyatomi@hosei.ac.jp

概要

大規模言語モデル（LLM）は自然言語処理で高い性能を示す一方，多様な文化的背景への対応には課題がある．本研究では，日本語コーパスで学習された LLM を活用することで，日本文化知識データベース NINJA を構築した．NINJA は，既存の文化知識データベース MANGO が持つ日本文化に関するデータと同数の 4,597 件生成し比較した結果，データの多様性を示す指標である Self-BLUE が 0.787 から 0.350 に改善した．さらに，日本の文化的背景への理解が必要な常識道徳推論タスクに対して，NINJA を外部情報として組み込んだ RAG（Retrieval Augmented Generation）システムを構築し，日本文化に即した推論における有用性を検証した．

1　はじめに

大規模言語モデル（Large Language Models; LLM）は急速に発展を遂げ，自然言語処理のさまざまな応用分野で優れた性能を示している．しかし，多様な文化的背景や価値観に即した応答を生成する能力については，いまだ多くの課題がある [1, 2, 3, 4]．特に OpenAI の GPT モデルは，主に英語を中心としたコーパスで学習され，さらにアノテーターの価値観や社会文化の影響を受けることから，特定の文化や社会的文脈に偏る可能性が指摘されている [5, 6]．こうした課題に対処するため，文化的な違いを単一のモデルに組み込み，多文化への対応を目指す取り組みが進められている一方で，各地域固有の文化や価値観に焦点を当てたモデルやデータの開発も行われている [7, 8, 9]．

日本においては，日本特有の常識や倫理観を言語モデルに反映させるための初の日本語常識道徳データセット JCommonsenseMorality（JCM）が構築され [9]，それを用いた日本特化型モデルの開発が検討されてきた [10, 11]．しかし，JCM では日本の文化を網羅するには限界があり，より広範かつ多様な常識や文化的背景をカバーするためには，さらに豊富なデータリソースの整備が求められる．

一方，人工知能分野において，人間の日常的な常識的知識（Commonsense Knowledge; CSK）を収集する研究は，長い歴史を持つ [12, 13, 14]．これらの研究の多くは，全世界共通の概念や文化に依存しない知識を扱ってきたが，近年では文化的背景を考慮した文化的常識知識（Cultural Commonsense Knowledge; CCSK）に焦点を当てた取り組みも進展している [15, 16, 17]．Nguyen ら [17] は，LLM を用いて，さまざまな言語圏から幅広い CCSK を収集する MANGO という手法とデータベースを提案した．MANGO は，11K 個の文化クラスターに対応した 167K 個の CCSK を収集しており，幅広い文化に対応したデータを大規模に収集できるという高い汎用性を持つ一方で，主に英語データで学習されたモデルである GPT-3.5 Turbo を使用しているため，各文化における網羅性には疑問が残る．特定の文化や社会背景をより深く捉えるには，対象とする地域に対応した言語で学習されたモデルから CCSK を収集する必要があると考えられる．

本研究では，日本語コーパスで学習された LLM を活用し，日本特有の文化的知識を収集した日本文化知識データベース **NINJA**（kNowledge dIstillation for Natural language processing - JApanese）を構築した．NINJA では，既存の CCSK データベースの概念をもとに，日本文化に関する短文を先行研究の MANGO と同数生成を行い，CCSK の生成に使用するモデルの違いが生成データの多様性に与える影響を調査した．また，NINJA の有用性を検証するため，日本の文化的背景への理解が必要なタスクを対象に，NINJA を外部情報として活用した RAG（Retrieval Augmented Generation）[18] システムを構築し，入力テキストに関連する CCSK を考慮して推論を行い，その推定能を評価した．

表 1　NINJA の構築手順と生成例

Step1: Extract only Japanese culture from MANGO and translate them into Japanese

MANGO（データ数: 167,396）　***Format: Concept — Culture — Assertion***

・Concept: music, Culture: Japan, Assertion: Japanese music culture combines traditional instruments with modern J-pop and K-pop influences.

・Concept: gift-giving, Culture: Japan, Assertion: Gifts in Japan are commonly given and received with two hands for respect and politeness.

MANGO$_{\text{JAPAN}}$（データ数: 4,597）

・Concept: 音楽, Culture: 日本, Assertion: 日本の音楽文化は、伝統的な楽器と現代的な J-pop や K-pop の影響が融合している。

・Concept: 贈り物, Culture: 日本, Assertion: 日本では、贈り物は敬意と礼儀のために両手で授受されるのが一般的である。

Step2: Generate assertions based on MANGO's concepts

NINJA$_{\text{-llmjp13B}}$（データ数: 4,597）

・Concept: 音楽, Culture: 日本, Assertion: 日本における伝統的な楽器の代表的な例として、雅楽で使われる和琴がある。和琴は十三弦の琴で、優雅な音色が特徴的である。

・Concept: 贈り物, Culture: 日本, Assertion: お中元やお歳暮など、季節ごとに決まった贈答品を贈る習慣がある。

NINJA$_{\text{-swallow8B}}$（データ数: 4,597）

・Concept: 音楽, Culture: 日本, Assertion: 歌舞伎の演目には、伝統的な日本音楽が使われる。

・Concept: 贈り物, Culture: 日本, Assertion: お祝いの席で、紅白の水引で飾ったのし袋に入れて渡す。

2　日本文化知識データベース

本研究では，日本語 LLM を用いて CCSK を生成・収集し，日本文化知識データベース **NINJA** を構築した．NINJA は，先行研究である MANGO を参考にしており，指定した概念（Concept）と文化（Culture）に基づいた文章（Assertion）を生成する．MANGO[1] では GPT-3.5 Turbo を用いて CCSK を生成しているが，特定の言語に特化した LLM を活用することで，より幅広く具体的な文化知識を抽出できると考えられる．そこで，NINJA では大規模な日本語コーパスで学習された LLM を採用し，MANGO と同数のデータを生成して比較を行い，生成に使用するモデルが CCSK の多様性に与える影響を検証した．

2.1　構築手順

まず，167,396 件の CCSK を含む MANGO から，Culture が ”Japan” に該当する 4,597 件のデータを抽出し，その Concept，Assertion をすべて日本語に翻訳した MANGO$_{\text{JAPAN}}$ を作成した．ここでは，言語間の翻訳を行うライブラリ deep-translator から GoogleTranslator を使用した．

次に，MANGO$_{\text{JAPAN}}$ から Concept を抽出し，それに関する 日本文化の Assertion を日本語 LLM で生成した．今回使用した日本語 LLM のモデルは，日本語と英語のデータを重点的に学習し，日本語に特化した LLM である LLM-jp 13B [19][2] と Llama-3.1-Swallow 8B[3] であり，ハイパーパラメータは Temperature を 1.0，Top-p を 0.95 に設定した．この手法で，各データ 1 件の Assertion を生成し，MANGO$_{\text{JAPAN}}$ と同数の新しいデータを収集した．

表 2　データの多様性に対する評価

	Self-BLUE ↓	distinct-1 ↑	distinct-2 ↑
MANGO$_{\text{JAPAN}}$	0.787	0.103	0.428
NINJA$_{\text{-llmjp13B}}$	0.430	0.195	0.763
NINJA$_{\text{-swallow8B}}$	**0.350**	**0.231**	**0.777**

2.2　生成データに対する評価

表 1 にデータの生成例を示す．Concept が「音楽」と「贈り物」の場合，MANGO$_{\text{JAPAN}}$ および NINJA の両方で日本文化を示す Assertion が生成されている．特に，NINJA の生成例では，「音楽」は和琴や歌舞伎，「贈り物」はお中元やお歳暮，水引やのし袋といった具体的な事例が生成されている．

NINJA が MANGO$_{\text{JAPAN}}$ よりも幅広く日本特有の文化を捉えているかを測定するため，生成データの多様性を評価した．評価指標として，総トークン数に対するユニークトークン数の比率を示す distinct-N [20] と，あるテキストとそれ以外のテキストとの類似度を示す BLUE [21] を全データに対して算出し平均した Self-BLEU [22] を用いた．distinct-N は値が高いほど幅広いトークンが生成できていることを示し，Self-BLEU は値が低いほどテキスト間の類似度が低く，多様なテキストを生成できているこ

1）　https://mango.mpi-inf.mpg.de/

2）　https://huggingface.co/llm-jp/llm-jp-3-13b-instruct

3）　https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1

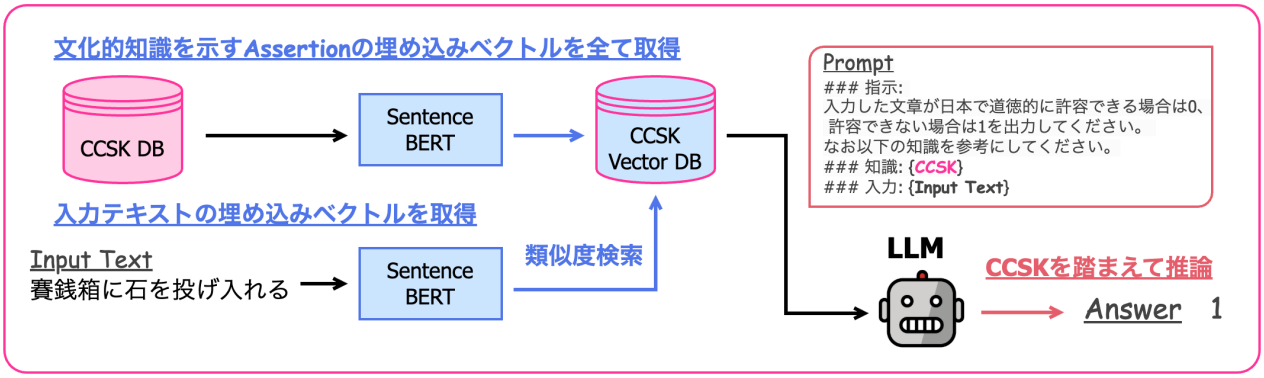

図 1　生成した CCSK（Cultural Commonsense Knowledge）を活用した RAG システム

とを示す．ここでは，各文書に対して Janome を用いて形態素解析を行い，名詞のトークンを抽出したあと，それぞれの指標を算出した．

表 2 から，NINJA は Self-BLUE の値が低い，つまりテキスト間の類似度が低く多様な文章を生成できていることを示した．また distinct-N においては， uni-gram および bi-gram の両方で NINJA が高い，つまりデータベースの中でより異なる単語が多く含まれていることを示した．これらの結果から，$\text{MANGO}_{\text{JAPAN}}$ と Concept，Culture，およびデータ数において同じ条件で構築された NINJA の方が，より多様なデータ生成を実現できていると言える．NINJA を構築する際に使用した LLM-jp 13B や Llama-3.1-Swallow 8B と比較して，MANGO で使用された GPT-3.5 Turbo は非常にパラメータ数が大きいが，NINJA の方が各指標で良化していることから，特定地域の文化を抽出するにはそれに対応した言語のモデルを用いることが有効だと示唆される．

3　実験

3.1　実験方法

我々は，NINJA の有用性を検証するために，言語モデルの日本における倫理的判断能力を評価するデータセット JCommonsenseMorality（JCM）[9][4] に対して，生成された CCSK を活用して推論を行う．JCM は，日本の常識道徳を反映させた初めてのデータセットであり，文章に対して道徳的に許容できるか否かの 2 種類のラベルが付与されている．このデータセットには，日本の文化や規範に関する知識がないと正確な判断が難しいケースがあり，文化知識を追加情報として付与することで，より適切な判断を促すことができると考えられる．

図 1 に，本研究で用いる日本の常識道徳推論を行う CCSK を活用した RAG システムを示す．まず，事前に CCSK データベース内の全データを埋め込みベクトルに変換した．次に，入力テキストの埋め込みベクトルとデータベース内の埋め込みベクトルとのコサイン類似度を計算し，データベースから最も関連性の高い CCSK を取得した．その後，類似度検索で取得した CCSK を追加したプロンプトをもとに LLM による回答生成を行い，ラベル推定をした．

3.2　実験設定

モデル　本実験で使用する常識道徳推論システムの中で行う，入力テキストに関連する CCSK を検索するための埋め込みベクトルの取得には，Sentence BERT [23][5] を用いた．また，推論に用いる LLM には，LLM-jp {3.7B, 13B} [19], ELYZA-japanese-Llama-2 {7B, 13B}[6]，Swallow {7B, 13B, 70B} [24]，Llama-3.1-Swallow {8B, 70B}，GPT-4o mini[7]（2024 年 7 月 18 日時点のモデル）を用いた．ハイパーパラメータは Temperature を 0.2，Top-p を 0.95 に設定した．

比較手法　ラベル推定には以下の 4 つの方法で行い，正解率（Accuracy）と F1 スコアで評価を行った．

1. Zero-shot
2. w/ $\text{MANGO}_{\text{JAPAN}}$ CCSK
3. w/ $\text{NINJA}_{\text{-llmjp13B}}$ CCSK
4. w/ $\text{NINJA}_{\text{-swallow8B}}$ CCSK

2, 3, 4 は，図 1 の RAG システムに外部情報として，それぞれの CCSK データベースを活用した．

4）https://github.com/Language-Media-Lab/commonsense-moral-ja

5）https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

6）https://huggingface.co/elyza

7）https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

表 3　JCM における CCSK を活用した各 LLM の評価結果

	Zero-shot		w/ $\text{MANGO}_{\text{JAPAN}}$		w/ $\text{NINJA}_{\text{-llmjp13B}}$		w/ $\text{NINJA}_{\text{-swallow8B}}$	
Model	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
LLM-jp-3-3.7b-instruct	0.310	0.076	**0.390**	0.039	0.342	0.080	0.321	**0.085**
LLM-jp-3-13b-instruct	0.810	0.756	**0.832**	**0.796**	0.821	0.781	0.823	0.780
ELYZA-japanese-Llama-2-7b-instruct	0.571	0.204	**0.609**	0.405	0.605	0.417	0.606	**0.429**
ELYZA-japanese-Llama-2-13b-instruct	0.615	0.506	0.596	0.339	0.624	0.528	**0.634**	**0.556**
Swallow-7b-instruct-hf	0.498	0.116	0.494	0.420	0.506	0.469	**0.511**	**0.481**
Swallow-13b-instruct-hf	0.603	0.362	0.610	0.406	0.615	0.432	**0.623**	**0.445**
Swallow-70b-instruct-hf	0.819	0.779	0.827	0.797	0.847	0.832	**0.852**	**0.837**
Llama-3.1-Swallow-8B-v0.1	**0.879**	0.869	0.876	0.866	0.866	0.859	0.878	**0.870**
Llama-3.1-Swallow-70B-v0.1	**0.929**	**0.924**	0.908	0.904	0.906	0.903	0.909	0.906
GPT-4o mini	**0.908**	**0.901**	0.899	0.894	0.894	0.889	0.902	0.897

評価用データ　評価には，JCM のテストデータ 3,992 件を用いた．割り当てられているラベルの内訳は，「道徳的に許容できる」が 2,124 件，「道徳的に許容できない」が 1,868 件である．

4　結果と考察

4.1　CCSK による効果

表 3 の結果から，LLM-jp や ELYZA LLM，Swallow では，プロンプトに CCSK を含めることで推定能が向上し，Llama-3.1-Swallow や GPT-4o mini では，推定能が変わらない，もしくは低下する結果となった．これらの違いは，モデルの Zero-shot における推定能に関連していると考えられる．Zero-shot でのスコアが低いモデルほど，CCSK の効果が顕著に現れ，追加情報として CCSK を有効に活用できている．一方で，Zero-shot でのスコアが高いモデルでは，CCSK が逆効果となる傾向があり，この原因として CCSK が限られた 4,597 件のデータソースに基づいていることが挙げられる．本実験では，入力テキストに類似したケースがデータベースに存在しない場合でも 1 つの CCSK を付加しているため，カバーできない事例では CCSK がノイズとなる可能性がある．また，生成された CCSK に対する正確性や文章の自然さの検証が行われていないため，今後大規模な CCSK データベースを構築する際にはこの点を考慮する必要がある．

4.2　CCSK の生成に用いるモデルの違い

$\text{MANGO}_{\text{JAPAN}}$ と NINJA（LLM-jp 13B，Llama-3.1-Swallow 8B）の比較を行った．CCSK を活用することで推定能が向上したモデル（LLM-jp, ELYZA LLM, Swallow）に注目すると，LLM-jp 13B を除き，NINJA が $\text{MANGO}_{\text{JAPAN}}$ を上回る結果を示した．その中でも高いスコアが出ている Swallow 70B では，$\text{NINJA}_{\text{-swallow8B}}$ が Zero-shot の F1 スコアを 5.8 ポイント向上させ，$\text{MANGO}_{\text{JAPAN}}$ より 4.0 ポイント高いスコアを達成した．この結果は，表 2 に示すように CCSK データベースが多様で具体的な文章を生成できることが要因だと考えられ，本タスクにおいても優れた効果が確認された．

さらに，$\text{NINJA}_{\text{-llmjp13B}}$ と $\text{NINJA}_{\text{-swallow8B}}$ を比較したところ，ほとんどのモデルで $\text{NINJA}_{\text{-swallow8B}}$ が高いスコアを記録した．NINJA の構築に使用したモデルの Zero-shot の F1 スコアも，Llama-3.1-Swallow 8B が LLM-jp 13B を 11.3 ポイント上回った．これらの結果から，CCSK 作成に使用するモデルには，日本の常識や文化を反映したモデルを採用することが重要であると考えられる．

5　おわりに

我々は，日本語 LLM を用いて日本文化知識データベースである NINJA を構築し，先行研究である GPT モデルを用いて構築された MANGO と比較した結果，日本文化に関する多様な CCSK の生成ができることを確認した．また，CCSK の有用性を検証するため，日本の文化的背景への理解が必要である JCM において CCSK を活用したところ，Zero-shot での推定能が低いモデルでは追加情報として CCSK の効果が見られたが，推定能が高いモデルにはノイズになってしまい，性能が低下した．

今後は，より大規模に特定の国特有の文化知識を収集する手法を検討し，生成テキストの多様性と正確性を担保したデータベースの構築を目指す．

参考文献

[1] Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACCL)**, pp. 5668–5680, 2024.

[2] Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 16366–16393, 2024.

[3] Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. Massively multi-cultural knowledge acquisition & lm benchmarking. **arXiv preprint arXiv:2402.09369**, 2024.

[4] Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. CultureLLM: Incorporating Cultural Differences into Large Language Models. In **Proceedings of 38th Annual Conference on Neural Information Processing Systems (NeurIPS)**, 2024.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023.

[6] Partha Pratim Ray. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. **Internet of Things and Cyber-Physical Systems**, 2023.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with Shared Human Values. In **Proceedings of the International Conference on Learning Representations (ICLR)**, 2021.

[8] Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. HAE-RAE Bench: Evaluation of Korean Knowledge in Language Models. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)**, pp. 7993–8007, 2024.

[9] 竹下昌志, ジェプカラファウ, 荒木健治. JCommonsenseMorality: 常識道徳の理解度評価用日本語データセット. 言語処理学会第 29 回年次大会, pp. 357–362, 2023. in Japanese.

[10] Yuu Jinnai. Does Cross-Cultural Alignment Change the Commonsense Morality of Language Models? In **Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP**, pp. 48–64, 2024.

[11] Takumi Ohashi, Tsubasa Nakagawa, and Hitoshi Iyatomi. Extended Japanese Commonsense Morality Dataset with Masked Token and Label Enhancement. In **Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)**, p. 3964–3968, 2024.

[12] Douglas B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. **Commun. ACM**, Vol. 38, No. 11, p. 33–38, 1995.

[13] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**, p. 4444–4451, 2017.

[14] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models. In **Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACCL)**, pp. 4602–4625, 2022.

[15] Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In **Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)**, pp. 67–78, 2022.

[16] Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting Cultural Commonsense Knowledge at Scale. In **Proceedings of the ACM Web Conference 2023 (WWW)**, p. 1907–1917, 2023.

[17] Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. Cultural Commonsense Knowledge for Intercultural Dialogues. In **Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)**, p. 1774–1784, 2024.

[18] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In **Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)**, pp. 9459–9474, 2020.

[19] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. **CoRR**, 2024.

[20] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A Diversity-Promoting Objective Function for Neural Conversation Models. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACCL)**, pp. 110–119, 2016.

[21] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)**, pp. 311–318, 2002.

[22] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A Benchmarking Platform for Text Generation Models. In **Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval**, pp. 1097–1100, 2018.

[23] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3982–3992, 2019.

[24] 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 **Proceedings of the First Conference on Language Modeling (COLM)**, 2024.