Doppelganger-JC：日中同形異義語の理解能力を測る LLM ベンチマーク

北村 優佳 [1,3]　黄 嘉皓 [1]　相澤 彰子 [1,2,3]
[1] 東京大学 大学院情報理工学系研究科　[2] 国立情報学研究所
[3] 国立情報学研究所 大規模言語モデル研究開発センター
{ykitamura,aizawa}@nii.ac.jp　jiahao-huang@g.ecc.u-tokyo.ac.jp

概要

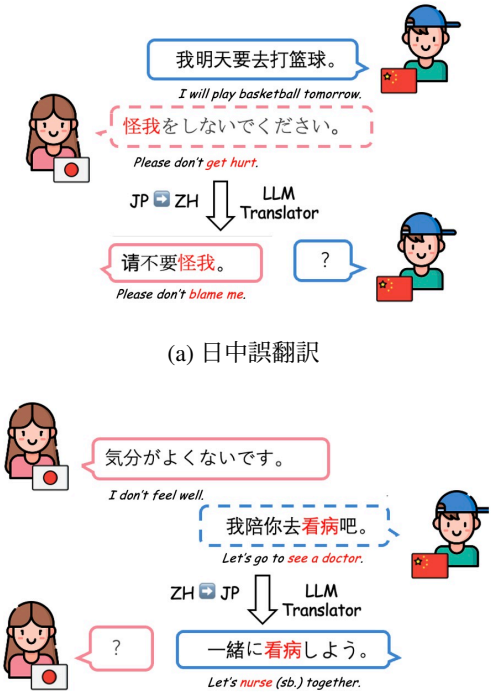
LLM の発達は目覚ましいが、依然として言語を跨ぐ同形異義語を扱うには課題がある。本研究では日中同形異義語 (表記は同一だが各言語で意味が異なる単語) に注目し、LLM がそれらを正しく扱えるのかを評価するためのベンチマークデータセット Doppelganger-JC を構築し、分析を行った。分析の結果、言語モデルが同形異義語を「理解しやすい」言語で解釈する傾向があることがわかり、我々はこれを homograph shortcut と名付けた。この傾向は、日中同形異義語の品詞が両言語で一致している場合に顕著にみられる。本データセットは、以下のリンクで公開している[1)]：
https://github.com/0017-alt/Doppelganger-JC

1　はじめに

近年 LLM の発展は目覚ましいが、英語の学習データが豊富であることから、LLM は多くの場合英語中心的であり、多言語対応が重要な課題になっている [1]。[2] によれば、LLM の 2 言語に跨る同形異義語の扱いにはまだ課題がある。

中国語と日本語は歴史的に互いに影響を与え合っており、両言語には多くの共通した単語が存在する。それらは両言語で似たような意味を持つ場合が多いが、意味変化を経て全く異なる意味を持つようになったものもあり、それらを同形異義語と呼ぶ。統計的には、日本語の頻出上位 2 万語のうち、50%は漢語、29%は同形語、6%は同形異義語とされる [3]。

図 1 に日中同形異義語の誤用例を示す。現状、LLM が日中同形異義語を扱う能力を体系的に評価するデータセットは我々の知る限り存在していない。そこで我々は (1) **Doppelganger-JC** というデータセットを構築し、(2) 日中同形異義語に関する LLM の性能評価と誤用の原因の分析を行った。



(a) 日中誤翻訳

(b) 中日誤翻訳

図 1: 日中両言語における同形異義語の誤用例

Doppelganger-JC には (1) 語義判定、(2) 文脈内語義判別、(3) 翻訳の 3 つのタスクがあり、それぞれ多肢選択問題として与えられる。評価に用いたモデルは、2 つの日本産 LLM、2 つの中国産 LLM と 3 つのその他の LLM である (表 1)。

本研究の貢献は以下である：

1) 本研究の内容は、国際学会 IJCNLP-AACL 2025 に採択されたものに基づく。

2) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3
3) https://ai.meta.com/blog/meta-llama-3/
4) https://telnyx.com/llm-library/mistral-7b-instruct-v0-2

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

モデル	事前学習コーパスの言語
llm-jp-3-7.2b-instruct3 [4]	英語 (950B) / 日本語 (592B)、韓国語 (0.3B)、中国語 (0.8B)[2]
Llama-3-ELYZA-JP-8B [5]	日本語 (追加事前学習), 主に英語だが、5%は 30 を超える言語を含むデータ (Llama-3-8B-Instruct[3])
Qwen2.5-7B-Instruct-1M [6, 7]	30 言語 (英語、中国語、スペイン語、フランス語、ドイツ語、アラビア語、ロシア語、韓国語、日本語、タイ語、ベトナム語など) [8]
Baichuan2-7B-Base [9]	公開されていないが、多言語対応していると述べられている
Llama-3.1-8B-Instruct [10]	主に英語だが、8%は 176 を超える言語を含むデータ
gemma-7b [11]	主に英語
Mistral-7B-Instruct-v0.2 [12]	公開されていないが、英語と Hinglish に対応している[4]

表 1: 実験で使用した LLM

- Doppelganger-JC という新しいデータセットを構築し、LLM が日中同形異義語を正しく扱う能力の測定を可能にした。
- Doppelganger-JC を用いた実験を通して、homograph shortcut (日中同形異義語を自身にとって「理解しやすい」言語で解釈する現象) が多くのモデルで発生することを指摘した。
- 言語学的特徴を踏まえた分析から、日中同形異義語が共通の品詞を持つ場合に homograph shortcut が起こりやすいことを示した。

2 関連研究

同形異義語 [13] は同形異義語を *"words in different languages share the same orthographic form"* と定義している。言語モデルがどのように同形異義語を扱うかについての研究もいくつか存在しており、[2] は英語-スペイン語、フランス語-ドイツ語単語ペアについて曖昧性解消、意味判定、意味制約タスクを行い、同形異義語に関して LLM の性能がランダムを下回ると報告した。[14] はドイツ語と英語のコードスイッチングタスクを行い、ニューラル LM が同形異義語の識別性能が最も良いことを示した。また、[15] は同形語ペアが言語間で異なる意味を持つかを判定するタスクを行い、BERT 単語埋め込みの類似度と並列コーパス中での共起率を用いることで識別性能が上がることを示した。

しかし、LLM が同形異義語の意味を正しく識別できているのかに注目したデータセットは我々の確認した限り存在していない。そこで我々は新たなデータセットを構築した。

日中同形異義語 日本はかつては無文字社会であったが、約 1,600 年前に漢字が中国から導入された [16]。日本語が文字を受け入れる中で、中国から入ってきた外来語が一般化・抽象化・特殊化などを経て、元の中国語の意味とは異なる意味を持つようになった [17]。

このような過程は他の言語ペアでも起こりうる。日中同形異義語に関するデータセットを構築することは、その手法を他の言語ペアにも適用できるという点で有用である。

3 データセット構築

Doppelganger-JC の構築は、(1) 単語-意味ペア構築、(2) 例文生成、(3) タスク生成、の流れで行った。以下に詳細を記す。

3.1 単語-意味セット構築

はじめに、日中対照漢字語データベース Version 2.00 (JKVC) [3] から日中同形異義語の収集を行った。JKVC からは以下の 4 タイプの日本語の単語のみを集めた；**Type-1**：日中の意味が全く異なる、**Type-2**：共通の意味と、それぞれ独自の意味がある、**Type-3**：日本語のみ独自の意味を持つ、**Type-4**：中国語のみ独自の意味を持つ。

その後、日本語話者が全単語が日本語で書かれた意味とペアになっているかを確認し、GPT-4.1[5] を用いてすべて中国語に翻訳した。その後、中国語話者が翻訳が不自然なものを修正した。

最終的な単語-意味セット数は 1,290 であり、各単語は日中共通義・日本語独自義・中国語独自義とセットになっている。各言語語の内訳は表 2 の通りである。

言語	Type-1	Type-2	Type-3	Type-4	合計
日本語	464	182	355	-	1,001
中国語	464	182	-	289	935

表 2: データセットの内訳

3.2 例文生成

LLM が異なる言語間で同形異義語を正しく扱えるかをより深く分析するため、収集した各単語についてそれぞれ 1 例文ずつ生成した。次段落で日本語例文の生成方法について説明するが、中国語に関しても同様である。

各同形異義語について日本語独自の意味を提示し、その意味を用いて GPT-4.1 にその単語を含む例

5) https://openai.com/index/gpt-4-1/

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

文を生成させた。また、同時にその例文の中国語・英語への翻訳、3 つの「誤った」中国語への翻訳も生成させた。この「誤った」例文のうち一つは、必ず同形異義語そのものを含むようにしている。この同形異義語を含むような翻訳を選択した場合、LLM は homograph shortcut を起こしているということができる。

生成文の質を保証するため、日本語話者・中国語話者が日中全ての生成文に対して (1) 流暢で自然か、(2) 例文に含まれる単語は指定した意味で使われているか、(3) 正しい翻訳文が同形異義語をそのまま含んでいないか、(4) 誤った翻訳文のうち少なくとも一つが同形異義語を含んでいるか、を確認した。

3.3 タスク設計

LLM の日中同形異義語処理性能を評価するために、語義判定タスク・文脈内語義判別タスク・翻訳タスクの 3 つのタスクを用意した。各タスクは 4 択問題として与えられる。以下日本語タスクについて説明するが、中国語に関しても同様である。

語義判定タスク LLM は語義を正しく選択する必要がある。プロンプトは日本語で書かれ、選択肢は日本語独自義、中国語独自義、無関係な義 2 つの 4 つで、この中から正しく日本語独自義を選ぶことになる。

文脈内語義判別タスク LLM は例文中の単語の意味を正しく選択する必要がある。プロンプトは日本語で書かれ、選択肢は語義判定タスクと同様である。

翻訳タスク LLM は与えられた中国語文の正しい日本語訳を選択することになる。プロンプトは日本語で書かれ、選択肢は正しい翻訳文と、生成した 3 つの「誤った」翻訳文である。

4 実験

4.1 実験設定

モデル 使用モデルは表 1 に示している。各モデルは自国の言語で、より性能が上がるという仮説のもと、複数の日本産・中国産モデルを採用している。また、比較のため、データセット中の各タイプから 50%の設問を取り出し、2 名の中国人留学生ボランティア (JLPT-N2 レベル[6]) に対して LLM と全く同じ指示を示して問題を解いてもらった。

6) https://www.jlpt.jp/e/about/index.html

評価指標 全タスクに関し、LLM に指示文とそれぞれの選択肢を与え、最も perplexity が低かったものの正答率を LLM の性能としている。

4.2 結果

表 3 と 4 は今回測定した各 LLM の性能である。人間はどのタスクについても高い性能を出しているが、LLM の性能は比較的低くとどまっている。特に、単語に関する文脈情報が不足しているため、語義判定タスクが LLM にとっては最も難易度が高くなっている。語義判定タスクでは正答率がランダム率よりも低い場合もあるが、文脈情報を与えられると (文脈内語義判別タスク) モデルの精度は向上する。

また、言語に注目すると、日本産モデルは日本語タスク、中国産モデルは中国語タスクでより良い性能を出すことがわかる。これは、各モデルの事前学習コーパス中の言語の偏りによるものだと考えられる。日中同形異義語に関しては、モデルは自身が「理解しやすい」言語により近いものをたとえ誤りであっても選ぶ傾向があると言える。この傾向を homograph shortcut と名付ける。

以下の章では、翻訳タスクに焦点を絞り、この homograph shortcut について分析を進める。

Model	語義判定		文脈内語義判別			翻訳		
	T-1	T-2	T-1	T-2	T-3	T-1	T-2	T-4
Human	95.04	93.40	93.40	94.47	92.20	95.28	95.82	93.82
llm-jp	**64.87**	**50.55**	**79.00**	49.17	**58.36**	65.49	**72.63**	73.08
ELYZA-JP	53.88	42.86	74.24	**51.38**	55.81	60.44	67.60	70.63
Qwen2.5-7B	30.60	25.27	59.09	44.20	27.20	64.18	70.39	**73.78**
Baichuan2-7B	20.91	18.68	57.14	44.20	46.74	58.90	62.57	64.34
Llama-3.1	41.16	34.62	67.75	45.30	46.74	**70.55**	71.51	72.38
gemma-7b	41.38	35.71	64.94	46.41	46.46	60.00	67.04	68.53
Mistral-7B	24.35	17.58	51.30	37.02	39.09	41.32	46.37	45.80

表 3: LLM の日本語タスク性能 (%)

Model	語義判定		文脈内語義判別			翻訳		
	T-1	T-2	T-1	T-2	T-4	T-1	T-2	T-3
Human	94.40	90.66	96.10	92.73	98.84	94.59	95.03	90.94
llm-jp	28.45	37.91	40.66	30.17	36.36	58.23	61.88	50.99
ELYZA-JP	24.14	29.67	47.47	35.75	50.70	52.81	58.56	48.16
Qwen2.5-7B	**62.93**	**54.40**	**72.97**	**48.04**	**53.85**	69.70	72.38	61.47
Baichuan2-7B	60.34	51.10	71.21	54.19	66.78	61.69	67.96	50.14
Llama-3.1	49.57	47.80	63.96	46.93	45.45	**72.08**	**76.24**	**66.57**
gemma-7b	53.66	51.65	56.70	39.66	47.20	61.04	65.75	54.39
Mistral-7B	44.83	42.86	53.85	42.46	47.90	49.78	56.35	42.78

表 4: LLM の中国語タスク性能 (%)

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

5 翻訳タスクに関する分析

5.1 Homograph Shortcut の割合

表 5 は全ての誤りのうち homograph shortcut を起こした割合、つまり選択肢の中から同形異義語そのものを含むものを選んだ割合を示している。この結果から、主な誤りの原因は homograph shortcut であるとわかり、homograph shortcut は LLM にとって大きな問題であることがわかる。

モデル	日-中	中-日
llm-jp	72.87	67.84
ELYZA-JP	84.42	74.53
Qwen2.5-7B	84.36	81.44
Baichuan2-7B	85.89	71.91
Llama-3.1	67.24	65.53
gemma-7b	81.14	78.25
Mistral-7B	68.23	75.10

表 5: 翻訳タスクにおいて、LLM が同形異義語そのものを含む選択肢を選んで誤った割合 (%)

5.2 同形異義語誤用の非対称性

図 2 は翻訳タスクにおける同形異義語の誤用タイプを示している。結果より、人間は 52.5%が日中双方向の誤りであるのに対し、LLM は誤りが双方向で起こる場合は限られており、日本産モデルは中日訳でより誤りを起こしやすく、中国産モデルは日中訳でより誤りを起こしやすいことがわかる。§ 4.2 と同様に、各モデルの事前学習コーパスの言語の偏りが原因で引き起こされていると考えられる。

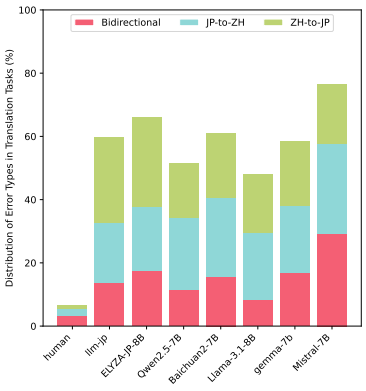


図 2: 翻訳タスクにおける同形異義語誤用の分布。青は日中訳のみの誤り、緑は中日訳のみの誤り、赤は両方向での誤りを示す。

5.3 日中同形異義語の品詞

データベース中の同形異義語を日中で品詞が同じ/異なるという 2 グループに分け、翻訳タスク性能を比較した (表 6)。品詞は、日本語については Mecab[7]、中国語については jieba[8]で取得した。

結果より、LLM は日中で品詞の異なる単語に対してはその用法や意味の違いをより捉えやすいことがわかる。依然として人間の精度には届かないものの、品詞のような言語学的な特徴を活用することは、LLM の同形異義語の誤用に関する問題の解決に近づく方法の一つだと考えられる。

Model	品詞が同じ		品詞が異なる	
	中-日	日-中	中-日	日-中
llm-jp	55.98	55.63	74.72	59.60
ELYZA-JP	52.45	50.07	72.24	55.77
Qwen2.5-7B	58.31	61.51	73.76	74.27
Baichuan2-7B	46.91	56.44	67.89	66.23
Llama-3.1	61.52	70.11	71.53	72.86
gemma-7b	54.00	54.31	73.54	64.99
Mistral-7B	28.92	45.80	49.53	54.42

表 6: Type-1, 2 の品詞が同じ/異なる単語それぞれの翻訳タスクの性能 (%)

6 結論

本研究では、LLM が日中同形異義語を扱う能力を測る Doppelganger-JC という新たなベンチマークデータセットを構築した。分析から、homograph shortcut という現象を発見し、LLM が同形異義語を自身の「理解しやすい」言語で解釈する傾向があることを指摘した。また、日中同形異義語が異なる品詞を持つ際は、文法制約からモデルは意味の違いを見分けやすく、homograph shortcut を回避しやすくなると示した。

これらの結果は LLM にとって日中同形異義語の扱いは未だ課題であることを示し、高品質データに加え品詞などの言語的特徴を活用した新たな手法を発展させる可能性を示唆している。

今後の課題として、本研究ではカバーしきれなかった日中同形異義語への対応、人手評価の大規模化などが挙げられる。また、多肢選択問題だけでなくより多様な方法での性能評価や、この手法の他言語ペアへの適用も考えられる。

謝辞 本研究の遂行にあたり、国立国語研究所の松下達彦先生に日中対照漢字語データベースを提供

7) https://taku910.github.io/mecab/
8) https://github.com/fxsjy/jieba

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

いただきました。また、松下先生並びに木下瞳氏には、多大なご助言、ご協力頂きました。ここに感謝の意を表します。

参考文献

[1]Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. **arXiv preprint arXiv:2410.05873**, 2024.

[2]Eshaan Tanwar, Gayatri Oke, and Tanmoy Chakraborty. Multilingual llms struggle to link orthography and semantics in bilingual word processing. **arXiv preprint arXiv:2501.09127**, 2025.

[3]松下達彦, 陳夢夏, 王雪竹, 陳林柯. 日中対照漢字語データベースの開発と応用. 日本語教育, Vol. 177, pp. 62–76, 2020.

[4]LLM-jp. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms, 2024.

[5]Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[6]An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. Qwen2.5-1m technical report. **arXiv preprint arXiv:2501.15383**, 2025.

[7]Qwen Team. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens, January 2025.

[8]An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[9]Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. **arXiv preprint arXiv:2309.10305**, 2023.

[10]Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[11]Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. **arXiv preprint arXiv:2403.08295**, 2024.

[12]Albert Q Jiang, A Sablayrolles, A Mensch, C Bamford, D Singh Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. arxiv. **arXiv preprint arXiv:2310.06825**, Vol. 10, , 2023.

[13]Ton Dijkstra, Jonathan Grainger, and Walter J.B. van Heuven. Recognition of cognates and interlingual homographs: The neglected role of phonology. **Journal of Memory and Language**, Vol. 41, No. 4, pp. 496–518, 1999.

[14]Igor Sterner and Simone Teufel. Tongueswitcher: Fine-grained identification of german–english code-switching. Association for Computational Linguistics, 2023.

[15]Mitko Nikov, Žan Tomaž Šprajc, and Žan Bedrač. Cross-lingual false friend classification via llm-based vector embedding analysis. Proceedings of the 10th Student Computing Research Symposium (SCORES'24), 2024.

[16]沖森卓也. はじめて読む日本語の歴史: うつりゆく音韻・文字・語彙・文法. ベレ出版, 2010.

[17]中川正之. 漢語からみえる世界と世間: 日本語と中国語はどこでずれるか. 岩波書店, 2013.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).