視覚言語モデル PIXEL による漢字表記体系の分析と考察

LI ZHENMING[1] 嶋田和孝 [1]
[1] 九州工業大学
li.zhenming714@mail.kyutech.jp
shimada@ai.kyutech.ac.jp

概要

漢字は中国や日本をはじめとする東アジア文化圏で広く用いられている文字体系であり、近年では視覚モデルの発展によりその文字形状を活用した知識転移の可能性が注目されている。本研究は漢字という文字システムに焦点を当て、中国語および日本語を対象とした異なる漢字表記に対して視覚モデル PIXEL を用いた学習と、その関連タスクへの転移効果を検討することを目的とする。まず、中国語の部首、簡体字、繁体字という三種類の表記を用いて事前学習済みの PIXEL モデルを追加学習する。その後、追加学習済みモデルを中国語簡体字感情分類、中国語攻撃的言語分類、日本語感情分類といった多様なタスクに適用し、性能評価を行う。これらの検証を通じて、漢字字形情報が転移学習に与える影響とその有効性を考察し、文字形状に基づく知識活用の可能性を明らかにする。

1 はじめに

自然言語処理における多言語モデルの発展に伴い、言語間もしくは文字表記間での知識転移は重要な研究課題として注目されている。従来のモデルは主として語彙ベースのトークン表現に依拠しており、文字体系の違いに対して柔軟性に限界がある。これに対して、PIXEL モデル [1] はテキストをレンダリング画像として入力し、視覚的表現を通じて言語表現を学習するという枠組みを採用している。PIXEL は画像上でのマスク復元学習により固定語彙を必要とせず、任意の文字表記に対応可能であり、視覚的類似性に基づく一般化や知識転移が期待される。特に、ある文字表記で獲得した知識を、視覚的に類似する他文字表記と関連タスクへ転移する応用は、実用・理論の両面で意義が大きい [2]。

漢字は中国、日本、台湾など東アジアの広範な地域で用いられる表意文字であり、一文字で広い意味を表現できるという特徴を持つ。漢字は通常複数の筆画から構成され、歴史的背景により地域ごとに異なる字体が発展してきた。中国大陸では画数の少ない簡体字が主に使用され、台湾・香港・マカオでは画数の多い繁体字が用いられている。一方、日本では伝来後に独自の発展を遂げ、日本独自の字体を含む漢字体系が形成された。これらの地域間には共通する字形も多く存在しており、視覚情報を活用するモデルにおいて、こうした共通字形を学習させることで、相互の効率的な知識転移が期待される。

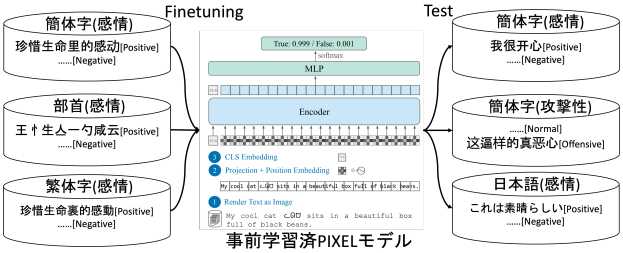

図 1 研究の全体像。PIXEL の画像は Rust ら [1] の論文の画像を引用

本研究では、図 1 のように、事前学習済みの PIXEL モデルに漢字の字形情報を追加学習し、追加学習済みのモデルが漢字関連タスクおよび関連言語への知識転移がどの程度可能かを検証し、その有用性と要素的要因について考察する。具体的には、中国語二値感情分類を基盤タスクとし、簡体字、繁体字、および漢字部首表記という三種類のデータで PIXEL モデルを学習させる。その後、中国語簡体字による感情分類、中国語簡体字による攻撃的言語分類、日本語感情分類の三つの評価タスクを通じて、中国語内でのタスク転移および日本語という異言語への汎用化能力を検証し、知見の整理と考察を行う。

2 関連研究

PIXEL は Vision Transformer (ViT) ベースモデルの代表例の一つである。元の PIXEL は英語コーパス

のみで事前学習されていたが，PIXEL-M4 [3] は英語，ヒンディー語，ウクライナ語，中国語（簡体字）の 4 言語という視覚的かつ言語的に多様なデータによる多言語事前学習へと拡張され，非ラテン文字言語における性能や文字体系をまたぐ転移能力において，モノリンガル版を大きく上回る成果を示した。Vision Transformer (ViT) は，Dosovitskiy ら [4] によって最初に提案されて以来，ViLT [5]，ALBEF [6]，PaLI [7] など，多数の後続研究に影響を与えている。

ViT の概念は，近年，マルチモーダル NLP タスクにも広く応用されている。Kim ら [5] は ViLT を提案し，畳み込みを用いない vision-and-language モデルとして，従来の視覚モデルよりも大幅に効率的でありながら，Visual Question Answering，画像–テキスト検索，視覚的推論において競争力のある性能を達成した。Ganz ら [8] は Question-Aware Vision Transformer (QA-ViT) を提示し，質問特化の認識を視覚エンコーダ内部に直接組み込むことで，クエリに応じた動的な視覚特徴適応を可能にし，多様なマルチモーダル推論タスクにおいて一貫した性能向上を実現した。

3　実験設定

3.1　研究のコンセプト

本研究では、中国語の二値感情分類（ポジティブとネガティブ）を基盤タスクとする。まず、中国語感情分類という同一タスクを前提に、簡体字、繁体字、部首の三種類の表記体系それぞれを用いて PIXEL モデルを学習させる。ある表記体系で学習したモデルは、その表記体系に基づく中国語の感情知識を保持しているとみなす。ここで、部首は漢字の構成要素であり画数が少なく、漢字の簡略化形とも捉えられる一方、繁体字は簡体字より画数が多く、より複雑な字形を持つ。したがって、部首、簡体字、繁体字はそれぞれ異なる水準の字形情報を提供し、PIXEL モデルに学習される字形表現の深さや特徴が異なることが想定される。

次に、学習済みモデルが獲得した知識を漢字関連タスクへどの程度転移可能かを検証するため、以下の三つのテストケースを設定する。

1. **中国語感情分類タスク:** 学習と評価を同一タスクで実施する。厳密には転移学習ではないが、各表記体系に基づき PIXEL モデルがどの程度の字形知識と感情判断能力を獲得したかを直感的に評価する。
2. **中国語攻撃的言語分類タスク:** 同じ中国語でありながら、対象現象が「感情」から「攻撃性」へと異なるため、タスク間転移として扱う。ここでは、各字形体系から学習された知識が、同一言語内で異なる自然言語処理タスクへ転移可能かを検証する。
3. **日本語感情分類タスク:** 日本語には漢字表記が含まれ、中国語漢字と部分的に共通性が存在する。一方で、ひらがな・カタカナといった中国語には存在しない表記体系も含まれるため、未知言語や未知表記を含む環境への知識転移を評価できる。

これら三つのテストケースは、いずれも漢字の字形表記と関連性を持ちつつ、(1) 同一タスク内評価、(2) 同一言語・異タスク転移、(3) 異言語転移という異なる観点からモデル性能を分析可能とし、PIXEL モデルにおける字形情報学習と転移能力に関する多面的な知見を得ることを目的とする。

3.2　実装設定

中国語簡体字のデータを基に、字形変換をする。Python のライブラリを使用して、中国語簡体字の文章を部首、繁体字に変換し、新たなデータを作る。中国語感情分類のデータを対象に、ラベルは変えずに、テキストのみに変換を行う。部首変換は Python の cjkradlib を使用し、RadicalFinder で候補構成要素を検索し、一番目の候補を選出する。繁体字変換は Python の opencc ライブラリを使用し、直接簡体字を繁体字に変換する。簡体字から直接部首や繁体字に変換する場合は、データの内容とラベルは変わらずに、ただ字形が変わるだけで、字形変形の影響がより明確に反映される。

モデルの選択はオリジナルの PIXEL モデルではなく、Kensen らの研究で 4 言語事前学習が拡張した PIXEL-M4 を使用する。オリジナルの PIXEL の事前学習は英語のみで、PIXEL-M4 は中国語簡体字を含めた。学習のパラメータは学習率を e-5、バッチサイズを 64、最大文長を 529 にする。

4　対象データ

本研究では複数の公開ソースからデータを引用する。まず、中国語の感情分類データはらの研究と公開された S2AP データセットを利用する。S2AP [9] は、Wan らによって発表された研究成果であり、中

国の SNS である Sina Weibo の投稿を対象に、大規模な感情分析を可能にするため構築されたデータセットおよび分析プラットフォームである。Wan らは S2AP プラットフォームを活用し、大量の Weibo 投稿を体系的に収集・前処理し、ノイズ除去した上で、ポジティブ／ネガティブ／ニュートラルといった感情ラベル付与を行った。本研究は上記のラベル付きデータを利用し、必要に応じて、ポジティブとネガティブの感情ラベルのデータのみ使用し、二値分類タスクにする。

次に、中国語の攻撃的言語データは COLD [10] データセットを使用する。COLD は、中国語攻撃的言語投稿のデータセットであり、Deng らによって公開された。約 3.7 万件の中国語コメントを収録し、「非攻撃的 (Non-Offensive) / 攻撃的 (Offensive) 」の二値ラベルが付与されている。

最後に、日本語の感情分類データは WRIME データセットを使用する。WRIME [11] は、日本語ソーシャルメディア投稿を対象とした感情強度推定データセットであり、Kajiwara らによって発表された研究成果である。約 40000 件の SNS 投稿に対し、投稿者自身が感じた主観的な感情強度（subjective） と第三者（読者）による客観的な感情強度（objective）を両方アノテーションした点が最大の特徴です。アノテーションは Plutchik の基本 8 感情（喜び／悲しみ／期待／驚き／怒り／恐れ／嫌悪／信頼）について 4 段階の強度で付与され、主観と客観の感情評価のズレを分析するための資源として設計されている。本研究では二値のラベルに統一するために、喜び、期待、信頼をポジティブに、残りをネガティブにまとめる。

上記の三つのデータセットから 20000 件を各データセットから抽出する。1 対 1 のラベル均衡を維持するよう、S2AP と WRIME からポジティブ 10000 件、ネガティブ 10000 件を抽出し、COLD から非攻撃的 10000 件、攻撃的 10000 件で抽出を行う。各データセットから抽出された 20000 件を 80% : 20% の割合で学習とテストに分割する。

5 結果と分析

学習のランダム性を考慮するため、各実験は 3 回実施し、Macro Average F1 の平均値を基に結果を評価した。実験の結果を表 1 に示す。例えば、表の 1 行目は、中国語簡体字の感情分類データで学習し、中国語簡体字の感情分類データで検証した 3 回の平均値が 0.774 であったことを意味する。表の中でまず学習のデータは簡 (情) が S2AP の学習データで、これを元に、部首変換したものが部 (情)、繁体字変換にしたものが繁 (情) となる。簡 (攻) は COLD の学習セット、日 (情) は WRIME の学習セットである。そして、テストのケースは簡情は S2AP のテストデータ、簡攻は COLD のテストデータ、日情は WRIME のテストデータである。

表 1 実験結果の Macro Average F1 値の三回の平均値。略称の意味は“簡”は中国語簡体字、“部”は中国語部首、“繁”は中国語繁体字、“日”は日本語。そして文字表記につく“情”は感情分類、“攻”は攻撃的言語分類を意味する

テスト	学習	F1 値
簡情	簡 (情)	0.774
簡情	部 (情)	0.557
簡情	繁 (情)	0.726
簡攻	簡 (攻)	0.737
簡攻	簡 (情)	0.498
簡攻	部 (情)	0.459
簡攻	繁 (情)	0.495
日情	日 (情)	0.682
日情	簡 (情)	0.493
日情	部 (情)	0.463
日情	繁 (情)	0.529

次に実験の結果について分析する。まず、部首文字表記を用いて学習した場合、オリジナルの簡体字で学習した場合と比較して大幅な性能低下が確認された。例えば、中国語感情分類テストにおいて、部首学習モデルの性能は 0.557 であり、簡体字学習モデルの 0.774 と比較して約 32%の低下が見られた。これは、部首変換が元となった漢字の字形情報の一部しか保持せず、視覚的情報量が著しく減少することにより、モデルが十分な特徴表現を獲得できなかったことが原因であると考えられる。一方で、繁体字による学習では性能低下は限定的であり、同タスクにおいて 0.726 と、オリジナル簡体字学習と比較して約 5%程度の差に留まった。繁体字は画数が多く字形構造が複雑であるものの、むしろ豊富な字形情報を保持することで、PIXEL の視覚処理機構が有効に機能したと推測される。以上より、PIXEL は字形の複雑化に対しては比較的堅牢である一方、情報量が大きく欠落する字形単純化には弱いということが示唆される。

次に、タスク転移の観点では、中国語攻撃的言語

データによる評価において、いずれの学習条件: 簡体字・部首・繁体字でも大幅な性能低下が確認された。中国語攻撃的言語を用いて直接学習した場合の性能 0.737 を基準とすると、感情分類データで学習したモデルはそれぞれ 0.498、0.459、0.495 に留まり、20%以上の性能低下が生じた。これは、同一言語・同一表記であっても、「感情」と「攻撃性」という異なる事象間には概念的乖離が存在し、視覚情報のみならず、タスクの相違が転移性能を制約することを示す結果である。

さらに、未知文字表記対応の観点では日本語感情分類データによる評価でも同様に性能低下が見られた。日本語感情分類データで直接学習した際の性能 0.682 に対し、中国語感情分類で学習したモデルは、簡体字 0.493、部首 0.463、繁体字 0.529 に留まった。日本語が漢字のみならず、ひらがな・カタカナといった異文字表記を含むことが未知視覚要素として作用し、全体性能を押し下げたと考えられる。しかしながら、このケースでは繁体字学習モデルの方が簡体字学習モデルよりも高い性能を示した点が注目される。これは、日本語漢字が構造的に繁体字に近い形態を保持しているものが多く、字形類似性がモデルの視覚表現転移を助けた可能性を示唆している。

日本語の検証ケースにおいて、繁体字で学習したモデルの性能が簡体字学習モデルよりも高いという現象をさらに分析する。追加考察では、繁体字の学習が日本語の漢字部分に寄与するか、それとも漢字以外の部分に寄与するかを分析することで、PIXEL の未知文字表記への対応メカニズムを解明する。

実験設定として、まず繁体字文字表記のみを対象とした中国語感情分類データを用い、PIXEL モデルを学習させる。次に、日本語感情分類データを検証データとして用いる際、日本語文中の文字を漢字とひらがな・カタカナに分離し、(1) 漢字のみを含む文字表記データと (2) 漢字以外（かなのみ）を含む文字表記データ の二種類を意図的に構成した。これらを用いて、繁体字学習済みモデルをそれぞれで評価することにより、繁体字学習効果が日本語の漢字部分に反映されているのか、あるいは非漢字部分にも及ぶのかを分析する。実験条件は前章の設定と同一とし、結果を表 2 に整理した。

その結果、日本語オリジナル文字表記での検証時の Macro-F1 は 0.529 であったのに対し、漢字のみを対象とした場合は 0.542 へと上昇し、逆に漢字以外のみを対象とした場合は 0.442 に低下した。この結果は、繁体字で学習された視覚表現が、日本語テキスト中の漢字に対して特に有効に機能していることを示している。すなわち、繁体字と日本語漢字の字形類似性が、視覚ベースモデルにおける転移学習の重要な要因として作用している可能性が高い。以上の追加検証により、PIXEL は未知の文字表記が混在しても、学習済みの字形知識を活用する能力を持つことが検証された。

表 2 繁体字変換後の文字表記で学習し、日本語のテストデータを漢字と漢字以外の部分に分けて、漢字のみと漢字以外で学習結果を検証する結果

テスト	F1 値
日本語文字表記	0.529
漢字のみ	0.542
漢字以外	0.442

6 結論

本研究では、PIXEL モデルが単一文字表記で学習した知識を、視覚的に近接した別文字表記および異なるタスクへ転移学習の効果を検証した。中国語簡体字の感情二値分類を基盤タスクとし、簡体字データに対して部首変換および繁体字変換を施すことで、ラベルを保持したまま視覚的に異なる三種類の文字表記データを構築する。そして、このデータを用いて PIXEL モデルを学習させたうえで、中国語簡体字の感情分類、中国語簡体字の攻撃的言語分類、日本語感情分類という三つのテストケースにより、文字表記間およびタスク間転移の効果を検証した。

本研究は、視覚ベース言語モデル PIXEL において、

1. 字形情報量の保持が学習性能に重要であること
2. 同一言語内であってもタスク差が転移性能を大きく制約すること
3. 未知文字表記が混在しても字形類似の活用は可能であること

を示した。今後は、複数文字表記環境における最適な視覚言語学習設計の検討が課題である。

謝辞

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2133.

参考文献

[1] Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. Language modelling with pixels. **ArXiv**, abs/2207.06991, 2022.

[2] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 16375–16387, 2022.

[3] Ilker Kesen, Jonas F. Lotz, Ingo Ziegler, Phillip Rust, and Desmond Elliott. Multilingual pretraining for pixel language models. **ArXiv**, abs/2505.21265, 2025.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. **ArXiv**, abs/2010.11929, 2020.

[5] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In **International conference on machine learning**, pages 5583–5594. PMLR, 2021.

[6] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. **Advances in neural information processing systems**, 34:9694–9705, 2021.

[7] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. **arXiv preprint arXiv:2209.06794**, 2022.

[8] Roy Ganz, Yair Kittenplon, Aviad Aberdam, Elad Ben Avraham, Oren Nuriel, Shai Mazor, and Ron Litman. Question aware vision transformer for multimodal reasoning. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pages 13861–13871, 2024.

[9] Shuo Wan, Bohan Li, Anman Zhang, Wenhuan Wang, and Donghai Guan. S2ap: Sequential senti-weibo analysis platform. In **Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part III**, page 745–749, Berlin, Heidelberg, 2020. Springer-Verlag.

[10] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A benchmark for Chinese offensive language detection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pages 11580–11599, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[11] Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. WRIME: A new dataset for emotional intensity estimation with subjective and objective annotations. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pages 2095–2104, Online, June 2021. Association for Computational Linguistics.