場面説明を活用したマルチモーダル漫画翻訳

横山泰知 [1]　戒能大翔 [1]　梶川怜恩 [1]　二宮崇 [1]　後藤功雄 [1]　石渡祥之佑 [2]　能地宏 [2]
[1] 愛媛大学　[2]Mantra 株式会社
{yokoyama@ai.cs., reon@ai.cs., ninomiya.takashi.mk@}ehime-u.ac.jp
goto.isao.fn@ehime-u.ac.jp,{ishiwatari,noji}@mantra.co.jp

概要

漫画の日英翻訳において，翻訳対象となるコマや吹き出し内のテキスト情報のみでは，翻訳に必要な情報が不十分である．本研究では，漫画画像に含まれる場面情報を文脈として活用し，翻訳精度の向上を図る漫画翻訳手法を提案する．漫画の日英翻訳タスクにおける評価実験の結果，従来手法と比較して，翻訳精度が改善することを確認した．

1　はじめに

日本の漫画は，日本国内のみならず世界中で親しまれている一方で，漫画の翻訳は人手で行われており，膨大な時間とコストを要する．そのため，ほとんどの作品は日本の国内市場から出ることはなく，第三者によって無許可に翻訳され海外で流通する被害が問題視されている [1]．正規版の国際的な普及のために，機械翻訳による迅速で高品質な漫画の自動翻訳が期待されている [2]．

漫画の機械翻訳では，吹き出しのテキスト単体だけでは，主語の省略などの文脈不足により，正確に翻訳するのに必要な情報が不足しているという課題がある [3]．図 1 に，実際に主語が省略された漫画のコマ画像を示す．このコマの吹き出しには，「きっとあの世で」，「やらなきゃよかったと思っているさ」とあるが，この吹き出しテキストには「誰が」あの世にいるのか，「誰が」後悔しているのかという主語が明示されていない．これにより，翻訳モデルは誤った主語を補完してしまい，正確な翻訳ができず，読者に本来の文意が伝わらない．先行研究では，これらの課題に対して，漫画のジャンルなどの属性情報 [4] や，画像から生成した作品の要約 [5] を文脈として入力することで，漫画の翻訳精度を向上している．しかし，漫画の属性や作品の要約では，誰が誰に対して発言しているかなどの主語補完に必要な細かな情報を得ることができない．

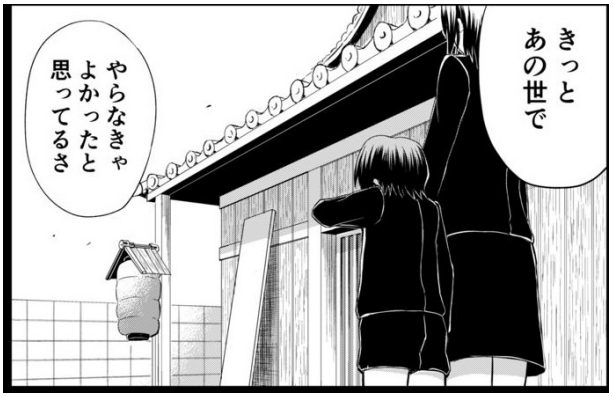

図 1　主語が省略された例 © 朽鷹みつき

これらの漫画の機械翻訳に関する課題に対処するために，本研究では，漫画のページやコマ画像の情報を考慮する手法を提案する．具体的には，ページ画像から，状況・登場人物・ストーリー・テーマといった，ページに描かれている場面の説明および曖昧性を解消する説明を作成し，翻訳時に活用する．漫画の日英翻訳タスクにおける評価実験の結果，漫画のページやコマの説明を考慮することで，漫画の翻訳精度が向上することを確認した．

2　関連研究

漫画翻訳におけるマルチモーダル情報の利用　漫画はテキストと画像が密接に関連し合うマルチモーダルであり，高品質な翻訳には画像情報の利用が不可欠である．Hinami ら [3] は，漫画翻訳を画像認識，文字認識，翻訳を含む包括的なタスクとして定義し，完全に自動化された翻訳パイプラインを提案した．彼らは，漫画画像から抽出した視覚特徴量を翻訳モデルに統合することで，テキストのみの翻訳に比べて性能向上することを示し，漫画特有の省略や曖昧性を解消するために，視覚情報が重要な役割を果たすことを示した．

翻訳における文脈情報の拡張と主語補完　日本語から英語への漫画翻訳において，主語の省略は深刻

な翻訳誤りを引き起こす要因である．この問題に対して Kaino ら [4] は，翻訳対象を吹き出しだけでなく，前後のコマのテキストも考慮することで翻訳精度を改善する手法を提案している．彼らは，文脈情報を拡張することで，省略された主語や指示語の推定精度が向上することを示した．しかし，テキスト情報のみに依存するため，視覚的な描写からしか推測できない情報は利用できない．

マルチモーダル LLM を用いた漫画翻訳　近年，大規模言語モデル（LLM）の発展に伴い，マルチモーダル LLM（MLLM）を用いた翻訳手法が提案されている．Lippman ら [5] は，MLLM を用いて，翻訳時に漫画画像とテキストを入力する手法を提案した．彼らは，ページ画像を MLLM に入力することで，文脈を考慮したより自然な翻訳が可能であることを示した．

3　提案手法

第 2 節で述べたように，漫画翻訳においては視覚情報の活用が不可欠である．特に，日本語から英語への翻訳では，主語を正しく補完するために，話者などの登場人物の文脈理解が求められる．本研究では，MLLM を用いて，画像からテキスト形式の場面説明を生成し，翻訳時の外部知識として利用する手法を提案する．場面説明は，ページ全体の文脈を捉えるためのページの場面説明と，各コマの局所的な状況を捉えるためのコマの場面説明を生成する．この 2 つの場面説明を翻訳時に吹き出しテキストと共に入力することで，主語補完に必要な情報を補い，翻訳精度の向上を目指す．

表 1 に、ページおよびコマの場面説明の生成で使用したプロンプトの例を示す．先行研究 [6] において，LLM による段落ごとの翻訳では，英語プロンプトが有効であると報告されている．この知見に基づき，本実験ではプロンプトは全て英語で記述した．

3.1　場面説明の生成

ページの場面説明　ページ単体の画像は，ストーリーの流れや登場人物の関係性を把握するため重要である．本手法では，翻訳対象の吹き出しがあるページ画像 1 枚を MLLM に入力し，場所や社会的情報を説明する Situation，ページ内に登場する人物に関する情報を説明する Characters，そのページで展開される出来事を説明する Story，ページ全体の雰囲気やテーマである Theme の 4 つを英語で出力する．

コマの場面説明　翻訳対象となる具体的な発話の文脈を捉えるため，コマ単位の説明文を生成する．入力するコマ画像は，データセットに付与された各コマ画像の座標情報をもとにページ画像を切り抜き使用する．さらに，対象のコマがどのような文脈の中に位置するかをモデルに認識させるように，ページの場面説明も同時に入力する．コマ画像 1 枚とそのコマが含まれるページの場面説明を MLLM に入力し，コマの場面説明を行うように指示し，英語で出力させる．

3.2　翻訳

最終的な翻訳では，上記のプロセスで生成したページの場面説明及びコマの場面説明を，2 パターンで翻訳対象の吹き出しテキストと組み合わせてモデルに入力する．具体的には、吹き出しテキストとページの場面説明を入力する手法と，吹き出しテキストとページの場面説明，コマの場面説明の 2 種類の場面説明を同時に入力する 2 パターンで翻訳を行った．これにより，吹き出しのテキストには含まれない主語補完や，曖昧性の解消を図る．

4　実験

4.1　データセット

本研究の評価実験では，研究目的で公開されている漫画の日英対訳データセットである Open-Mantra[1] [3] を使用した．本データセットは，5 つの異なる作品からなる計 5 巻，合計 214 ページ（1,593 文対）で構成されている．データセットに付与されたテキストボックスの位置や読み上げ順序から吹き出しと画像のペアを構築し，評価に用いた．

4.2　実験設定

本研究では，MLLM として，gpt-4o-2024-08-06[2] [7] を使用した．モデルの推論においては，生成結果の再現性を担保するために，温度パラメータを 0 とし，最大生成トークン数は 500 に設定した．本実験では，構造化出力[3]を用いて出力フォーマットを JSON 形式に設定した．

本実験では，翻訳時の吹き出しテキストの入力単

1） https://github.com/mantra-inc/open-mantra-dataset
2） https://platform.openai.com/docs/models/gpt-4o
3） https://platform.openai.com/docs/guides/structured-outputs

表 1　本研究において使用した場面説明と翻訳の指示例

タスク	例
ページの場面説明	I will give you a Japanese manga page. This manga page is copyright-free. Please describe the scene on this manga page in English. Follow the guidelines below for making the scene description. 1. situation: Information about the place and social situation. 2. characters: Information about the characters (name, gender, age, etc.). 3. story: Description of the scene depicted on the page. 4. thema: The thema on the page. For item 2 (characters), include the names of the characters in both Japanese and English. Additionally, write your response within square brackets[] in the format shown below: Output: [1. situation: 2. characters: 3. story: 4. theme:]
コマの翻訳	I will give you a description of a Japanese manga page. This manga is copyright-free. Here is a description of a manga page. {explain} I will give you an image of a manga panel on the manga page of this description. Please explain the situation of the panel in this page in English. Write your response within square brackets[]. Output: []

位は，吹き出しに含まれるテキスト単位，コマに含まれるテキスト単位，ページに含まれるテキスト単位の 3 種類で実験を行った．また，比較手法として漫画画像を一切使用せず吹き出しテキストのみで翻訳を行う手法，吹き出しテキストと翻訳対象となるテキストが含まれる漫画のページ画像を用いて翻訳を行う [5]2 つの手法を提案手法と比較する．吹き出しテキストのみで行う翻訳手法とその他の手法を比較することで，画像内の文脈情報が翻訳性能に与える影響を示し，吹き出しテキストと漫画のページ画像を用いた翻訳手法と提案手法を比較することで，文脈情報の活用方法が翻訳性能に与える影響を示す．

4.3　評価方法

生成された翻訳文の定量評価を行うため，前処理として出力テキストの正規化を行った．具体的には，モデル出力に含まれる不要な特殊記号や余分な空白を除去し，正解文及び出力文の双方に対して小文字化を適用した．

評価指標には，翻訳タスクで広く用いられる自動評価指標として，表層的な一致度を評価する BLEU[4)] [8] と，意味を考慮した評価を行う xCOMET[5)] [9] の 2 種類を採用した．

表 2　提案手法とベースラインの定量評価結果

手法	BLEU	xCOMET
吹き出し単位のテキスト	19.94	0.8328
+ ページ画像	**21.13**	0.8426
+ ページ場面説明	20.66	**0.8435**
+ ページ場面説明 + コマ場面説明	20.48	0.8427
コマ単位のテキスト	20.72	**0.8530**
+ ページ画像	**22.51**	0.8526
+ ページ場面説明	20.46	0.8491
+ ページ場面説明 + コマ場面説明	19.96	0.8493
直前のコマを含むコマ単位のテキスト	20.51	0.8364
+ ページ画像	19.48	0.7998
+ ページ場面説明	20.32	0.8242
+ ページ場面説明 + コマ場面説明	**20.76**	**0.8547**

4.4　実験結果

表 2 に各手法における，BLEU と xCOMET の評価結果を示す．それぞれの入力単位ごとに区切り線を挿入しており，「+」は吹き出しテキスト以外で翻訳時に入力した要素を示している．表中の太字は各

4）https://github.com/mjpost/sacrebleu

5）https://huggingface.co/Unbabel/XCOMET-XL

原文	「あさがお」の限定キーホルダー
参照訳	key chain of ″Asagao″. limited edition
テキストのみ	limited keychain of ″morning glory″
ページの場面説明	limited edition ″Asagao″ key holder

図 2　出力結果 © 朽鷹みつき

入力テキスト単位ごとの最高値，下線は全設定における最高値を示す．

全体として，テキストのみを入力した場合と比較して，画像や場面説明といった視覚情報を統合した場合の方が高いスコアを示す傾向が確認されたことから，視覚特徴量が漫画翻訳において性能向上に寄与することが確認された．

吹き出し単位のテキストの翻訳　ページ画像を入力した手法が最も高い BLEU を記録した．一方で，xCOMET においてはページの場面説明を入力した手法が最も高く，翻訳に必要な文脈情報を的確にモデルに伝達できている可能性を示している．

コマ単位のテキストの翻訳　ページ画像を入力した手法が BLEU，xCOMET において高い性能を示した．一方で，場面説明を用いた手法では，BLEU，xCOMET の両方が比較手法に劣っており，入力した場面説明がノイズとなり，モデルの推論精度を低下させたと考えられる．

直前のコマを含むコマ単位のテキストの翻訳　ページ画像を用いた手法では，BLEU，xCOMET の両方のスコアが急激に低下しており，全実験設定の中で最低値となった．一方で，ページ，コマの場面説明を入力した手法では，xCOMET が全実験設定で最高性能を達成し，BLEU においても入力単位内で最も高いスコアとなっており，提案手法の有効性が確認された．このことから，直前のコマを考慮したコマ単位などの複雑な文脈を扱う翻訳を行う場合，画像情報を直接入力するよりも，場面説明として構造化されたテキスト情報に変換して入力する方が有効であると考えられる．

4.5　事例分析

テキストとページの場面説明を入力して翻訳を行った結果と，テキストのみを入力して翻訳を行った結果を比較し，分析する．

図 2 に実際のコマの画像と入力した吹き出しテキストを示す．この作品では，「あさがお」はロケットの名前を指しており，参照訳では「Asagao」となっているが，テキストのみを入力した場合，花のあさがおの英名である「morning glory」と誤訳されている．これに対し，吹き出しテキストとページの場面説明を入力した場合では，場面説明として「holding up a key holder labeled "ASAGAO"」という記述があり（入力された場面説明全体は付録の図 3 参照），「Asagao」と正しく翻訳されている．このことから，提案手法では場面説明を文脈情報として活用できていることがわかる．

5　おわりに

本研究では，漫画の機械翻訳における，文脈情報不足による誤訳という課題に対し，ページ，コマ画像から生成した場面説明を文脈情報として翻訳に活用する手法を提案した．評価実験の結果，xCOMET では全実験設定で最高性能を達成し，提案手法の有用性を示した．一方で，入力単位によっては場面説明がノイズとなって性能が低下してしまうことも明らかとなった．今後の展望として，一貫した性能向上に向け，より長い文脈の考慮や，新たな文脈の活用法を探求する．

謝辞

本研究は国立研究開発法人情報通信研究機構の委託研究（課題番号：225）および JSPS 科研費 JP24K15071 の助成を受けたものです.

参考文献

[1] 日本経済新聞. 海賊版サイト　漫画、不正に「ただ読み」, 2026. https://www.nikkei.com/article/DGKKZO93549400U6A100C2EA2000/.

[2] 日本経済新聞. 漫画を AI で多言語翻訳、迅速輸出で海賊版を防止　文化庁が人材育成, 2026. https://www.nikkei.com/article/DGXZQOUD200JQ0Q5A221C2000000/.

[3] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards Fully Automated Manga Translation. **arXiv:2012.14271**, 2021.

[4] Hiroto Kaino, Soichiro Sugihara, Tomoyuki Kajiwara, Takashi Ninomiya, Joshua B. Tanner, and Shonosuke Ishiwatari. Utilizing Longer Context Than Speech Bubbles in Automated Manga Translation. In **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation**, pp. 17337–17342, 2024.

[5] Philip Lippmann, Konrad Skublicki, Joshua Tanner, Shonosuke Ishiwatari, and Jie Yang. Context-Informed Machine Translation of Manga Using Multimodal Large Language Models. **arXiv:2411.02589**, 2024.

[6] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting Large Language Model for Machine Translation: A Case Study. **arXiv:2301.07069**, 2023.

[7] OpenAI. GPT-4o System Card. **arXiv:2410.21276**, 2024.

[8] 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**, pp. 311–318, 2002.

[9] Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent Machine Translation Evaluation Through Fine-Grained Error Detection. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 979–995, 2024.

1. situation: The scene takes place outdoors on a grassy hill under an open sky. The setting suggests a casual and friendly social situation, possibly involving a discussion or exchange between the characters. 2. characters: The page features a young male, possibly a teenager or in his early twenties, wearing a short-sleeved shirt. Another character, a girl with long hair, is also present. Their names are not explicitly given here. An older man appears briefly, suggesting a friendly, possibly familial association with the young man. 3. story: The young male character is **holding up a key holder labeled "ASAGAO"** and mentions that it is something important that the girl had put in. He notes that it was sold 10 years ago, indicating its rarity or sentimental value. The girl seems curious and a bit flustered. The older man appears to be part of the conversation or situation, possibly supportive or sharing insights. 4. theme: The theme revolves around nostalgia, sentimentality, and the significance of seemingly small but cherished items. The exchange hints at understanding and fulfilling someone's interests or desires, showcasing connections and relationships.

図 3　図 2 の翻訳で使用した場面説明

付録

A　場面説明

図 2 のページの場面説明を活用した翻訳で使用した場面説明を図 3 に示す。場面説明の太字部分は，特に有効だと思われる箇所である．