動画キャプションモデルを用いた字幕翻訳の検討

成浦拓音　品川政太朗　須藤 克仁　中村 哲
奈良先端科学技術大学院大学
{nariura.takuto.nu7, sei.shinagawa, sudoh, s-nakamaura}@is.naist.jp

概要

機械翻訳において映像情報を用いて翻訳精度を改善する研究が行われている一方，映像情報を含むコーパスには限りがあり，大規模に学習することが難しい．本研究ではこのような問題を回避し，映像情報を話し言葉翻訳に活かす方法を提案する．本手法は映像情報を動画説明文に変換して取り扱うことで，事前学習済みの言語モデルが学習しやすい問題設定とし，規模の小さい視覚情報付きコーパスでも事前学習による性能向上が期待できる．VISA データセットを用いた実験の結果，多義語を含むデータセットにおいて動画説明文を用いて学習を行う事で，動画説明文を用いない場合より BLEU 値が 0.24 ポイント，METEOR 値が 1.19 ポイント上昇する結果を得られた．

1　はじめに

機械翻訳技術の発展により，実時間での同時翻訳や映画・漫画といった視覚情報を含む発話文や説明文を多言語に翻訳する研究が盛んに行われてきている [1, 2]．新聞や Web 上の文章をはじめとする文章の文脈や文法が意識される書き言葉とは異なり，同時翻訳や映画などの視覚情報を含む発話や会話には，指差しや視覚情報を共有しているという前提での主語の省略や，英語の “figure” やフランス語の “grille” といった場面に応じて異なった意味を含む多義語が事前の文脈情報なしに含まれることがある．

このような問題に対しては，視覚情報を考慮して翻訳することが有用だと考えられており，視覚情報を用いた翻訳手法の研究が進められている．実際に，Multi30k [3] や VATEX [4] など複数の視覚情報付きデータセットにおいて、視覚情報を追加することにより翻訳精度を改善できることが報告されている [5, 6]．

一方，映像情報を含む対訳コーパスの数は数千文から多くて数十万文と規模が小さく，大規模に学習することが難しい．

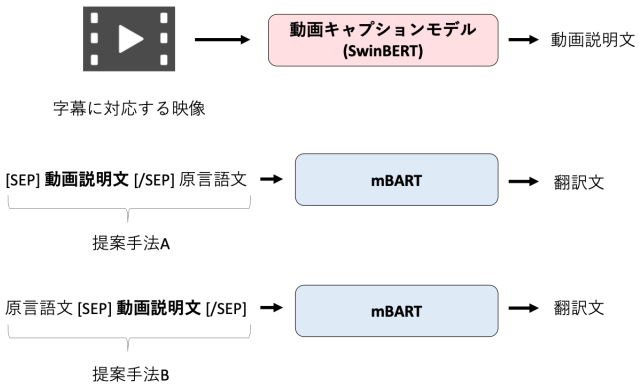

図 1　提案手法

本研究では視覚情報を動画説明文として受け取る事で話し言葉翻訳において改善が見込めると仮説を立て，大規模自然言語モデルにおいて視覚情報を取り扱う手法を提案する．本手法は視覚情報を事前に学習した動画キャプションモデルによって動画説明文へと変換し，変換した動画説明文を大規模自然言語モデルの入力として取り扱うことで事前学習済みの自然言語モデルを用いて学習を行う．自然言語処理では mBART [7] や T5 [8] といった大規模な事前学習を行った機械学習モデルを特定のタスクにファインチューニングすることが一般的であり，これらのモデルは数千万文以上からなる大規模なデータセットや対訳コーパスを用いて事前学習を行うことで事前知識を得ることを可能にしている．本提案手法を用いることで大規模データセットで事前学習された大規模自然言語モデルを直接的に変更することなく，そのまま活用することが可能となり，数千文から数十万文程度と規模が小さい視覚情報付き対訳コーパスでも事前学習による精度の向上を期待することができる．

提案手法を用いることで翻訳精度が向上することを明らかにするため，VISA データセット [9] を用いた実験を行った．実験の結果，多義語を含むデータ

セット (Polysemy) において動画説明文を用いた場合には，翻訳精度が改善する結果を得られた．

2 関連研究

2.1 動画翻訳

動画翻訳は映像情報を追加の入力として受け取ることで, 翻訳精度を改善することを目的とした研究である. 既存のデータセットとしては VATEX データセットや How2 データセット [10] が知られているが,VATEX データセットには映像情報を説明する動画説明文が付与されており，マルチモーダルな情報を活用することによる利点が十分に含まれていないという指摘も存在する [11]．Li らは OpenSubtitles データセット [12] から収集した日英字幕の対訳コーパスと，字幕情報に沿って取得した 10 秒間の映像情報を組み合わせた VISA データセットを構築した [9]．VISA データセットには字幕情報だけでは翻訳時の語義の曖昧性が解決できないとクラウドワーカーによって判断された文章のみが含まれており，マルチモーダルな情報を活用することによって語義の曖昧性を解決できることが期待されている．

映像情報を用いた機械翻訳モデルの多くは動画をフレーム単位に分けた後，事前学習済みの画像エンコーダに入力して画像特徴量を取得したり，モーションエンコーダを用いて動画特徴量を取得するなどの手法が一般的である [6, 13]．一方，本研究では視覚情報を直接翻訳モデルの入力として扱わないという点で既存研究とは異なる．

2.2 大規模言語モデル

ニューラルネットワークを用いた自然言語処理では事前学習を実施した大規模言語モデルをタスクの用途に合わせてファインチューニングをする手法が一般的であり，大量のコーパスを用いて事前学習を行うことでタスクによらない事前知識を獲得できることが知られている．機械翻訳でも大規模モデルを用いる手法が提案されており，Liu らは大規模な対訳コーパスを使用することで最先端の翻訳精度を実現する mBART を提案している．また Huang らは多義語の単語を説明する辞書情報を結合した入力を用いて，大規模事前学習モデルを微調整することでタスクを解く GlossBERT を提案している [14].

視覚情報付き翻訳においても大規模な事前学習済みモデルを使用する手法が提案されている．

表 1 VISA データセットの内訳

データセット名	Train	Val	Test
Omission (O)	17,160	1,000	1,000
Polysemy (P)	18,580	1,000	1,000
Combined (O + P)	35,740	2,000	2,000

Caglayan らは言語情報と視覚情報を踏まえた事前学習を行う手法である VTLM を提案した [5]．一方 Caglayan らの提案する事前学習手法はマスキングされた画像の分類問題をする追加タスクを実行する必要があり，それを行うためには事前に視覚情報の内容を物体検出した上でマスクを行えるように前処理をする必要がある．また映像情報のような時系列情報をそのまま結合することはできない．一方，本研究では，動画キャプションモデルで映像情報を事前に言語化することで，視覚情報を扱う難しさを回避しつつ，翻訳に有用な視覚情報を大規模言語モデルに自然に取り込める可能性が期待できる．

3 提案手法

本稿では動画説明文を用いることで翻訳に有用な視覚情報を大規模自然言語モデルに取り込める仮説を立て，画像特徴量や動画特徴量を翻訳モデルの入力として受け取る代わりに，大規模自然言語モデルの追加入力として視覚情報を説明する動画説明文を付与した上で翻訳を行うこととした．提案手法を図 1 に示す．VISA データセットにて各コーパスに紐づいている動画情報を動画キャプションモデルを使用して動画説明文へと変換し，それらを追加の入力文として取り扱う．動画の説明文は Huang らが提案した GlossBERT などの手法を踏まえて，入力として加える翻訳文に結合することとした．動画説明文は 2 つの特殊トークンである $[SEP]$ と $[/SEP]$ で挟んだ上で入力に結合する事とし，動画説明文の結合位置には原言語文の前後 2 通りが考えられるため，両方の場合を考慮して説明文を入力文の文頭に結合する場合 (提案手法 A) と文末に結合する場合の 2 通りで実験 (提案手法 B) を行った．動画キャプションモデルには最先端モデルの一つであり，映像を含む対訳コーパスである VATEX データセットで事前学習された SwinBERT [15] モデルを使用することとした．動画説明文の例を表 2 に示す．

表 2　動画説明文の例

動画ファイル名	説明文
omission_359950_17.mp4	a man is talking to the camera and then a man is talking.
polysemy_1204975_4.mp4	a group of people are sitting around a table and talking about poker.

表 3　提案手法 2 の評価値 (†: VMT [9] から引用)

	Omission			Polysemy			Combined		
手法	BLEU	METEOR	RIBES	BLEU	METEOR	RIBES	BLEU	METEOR	RIBES
VMT †	5.63	20.10	10.98	7.40	22.33	12.64	12.89	28.45	19.45
mBART	**9.97**	**32.14**	**20.77**	11.22	33.97	21.49	12.48	36.64	24.59
提案手法 A	9.67	31.73	20.48	10.34	34.28	21.53	13.05	**37.11**	25.17
提案手法 B	9.79	32.11	20.71	**11.46**	**35.16**	**22.21**	**13.21**	37.07	**25.46**

4　実験

事前学習済みの自然言語モデルに動画説明文を付与することで，翻訳精度が向上することを明らかにするために以下の実験を行った．

4.1　実験設定

本実験では映像情報付き翻訳データセットとして，VISA データセットを用いる．VISA データセットの内訳を表 1 に示す[1)]．VISA データセットには多義語の翻訳を含むデータセットである Polysemy と主語の省略を含むデータセットである Omission，その 2 つを組み合わせた Combined の 3 つの組み合わせが存在し，今回の実験では全てのデータセットの組み合わせで実験を行った．事前学習済み大規模自然言語モデルには CC-25 データセットで事前学習された mBART を用いた．なお，mBART の著者らが公開している mBART モデルとトークナイザには特殊トークン [*SEP*] と [/*SEP*] が含まれていないため，今回の実験では事前に使用する特殊トークンをトークナイザに追加しておき，特殊トークンに対応する単語埋め込み表現をランダムな初期値で初期化した上で学習を行った．

今回の実験では日英翻訳を行うため，日本語の原言語文に対して SwinBERT から出力された英語の説明文を結合して実験を行っている．また動画説明文を用いずに mBART を学習させた場合とも比較を行うこととした．テストデータに対する推論時にはビームサーチを用い，ビーム長は 10 とした．また実験コードの実装には fairseq [16] を用いた．

1)　映像が破損していて使用できなかったデータを除外した上での内訳

4.2　実験結果

実験結果は既存研究と同じく BLEU (n=4) [17] と METEOR [18]，RIBES [19] を用いて評価した．VISA データセットの文章は短い会話文が多いため，BLEU 計算時には文章の長さによる影響を緩和する手法として，既存研究と同じように Chen らの Smoothing 4 method を用いている [20]．それぞれのデータセットでの実験結果を表 3 に示す．全体として動画説明文を文頭に付与する場合よりも動画説明文を文末に付与する方が良い結果を収めている．また文末に動画説明文を付与した場合には動画説明文を付与しない場合よりも，幾つかの指標で最も良い結果を収めた．特に Polysemy データセットでは全ての指標において他のモデルを上回る結果を得ることができている．

動画説明文が有効に働いたと考えられる Combined データセットにおける出力例を表 4 に記載する．たとえば「シャワーを浴びようとしてた」という原言語文に対しては mBART は “He” と「彼」を主語にしているのに対して，提案手法 B は “I” と自分を主語とした発話として翻訳することが出来ている．また「気をつけろ」という原言語文に対しては mBART は「体に気をつける」といった意味合いである “Take care of” が出力されているのに対して，提案手法 B では「注意する」という意味合いを持つ “Be careful” が出力されており，動画情報を踏まえることによって複数の意味を持つ多義語を場面に応じて使い分けられている．また「体育館に戻られた方が良いと思います」という原言語文に対しては，体

育館を表す “gym” という単語を適切に出力できており，事前知識を活用して翻訳を行えている．

一方，動画説明文が有効ではなかったと考えられる出力例を表 5 に記載する．「金を待っているだけだ」という原言語文では，どの手法でも省略された主語を間違える結果となっている．日本語の会話文では場面に応じて主語が省略されることがあるが，今回のデータセットには話者情報が含まれていないため状況を説明する説明文だけでは対象の発話が誰に対する発話であるかを判別することが難しく，省略された主語を推測することに失敗したのだと考えられる．このような問題に対しては映像情報だけを用いるのではなく，話者情報を用いたり，音声情報や顔の表情から感情を推論するといったマルチモーダルな手法を活用することで更に翻訳精度を改善できるのではないかと考えている．

5 まとめと今後の展望

本稿では動画情報を動画説明文に変換し，大規模自然言語モデルの追加入力として取り扱う手法について検討を行った．VISA データセットを用いた実験の結果，いくつかの指標とデータセットにおいて提案手法は動画説明文を用いない場合よりも良い翻訳精度を実現することができた．今後の展望としては物体検出を行った上で，それぞれの物体に対して個別にキャプションを付与することで更に細かな動画説明文を取得することや，映像情報から話者情報や話し手の表情・声の調子や抑揚といった話者に着目したマルチモーダルな情報についても取得・活用することを検討している．また翻訳文章が長く複雑であったり，主語が推測しにくいような文章の場合にも本手法が今回の実験と同様に翻訳精度を向上するかどうかについても検討したいと考えている．

表 4 説明文が有効に働いたと考えられる出力例

原言語文	気をつけろ
リファレンス	So just be careful, OK?
mBART	Take care of yourself.
提案手法 A	Be careful.
提案手法 B	Be careful.
VMT	Be careful.
原言語文	シャワーを浴びようとしてた
リファレンス	I was about to take a shower.
mBART	He was trying to get a bath.
提案手法 A	I was trying to get a bath.
提案手法 B	I was trying to get a bath.
VMT	He was trying to get the shower.
原言語文	体育館に戻られたほうが良いと思います
リファレンス	I think you better head back to the gym.
mBART	I think we should go back to the gym.
提案手法 A	I think you should go back to the gym.
提案手法 B	I think you should go back to the gym.
VMT	I know it’s time to get the didn’t notice how long down.

表 5 説明文が有効に働かなかったと考えられる出力例

原言語文	金を待ってるだけだ
リファレンス	They’re expecting cash.
mBART	I’m just waiting for the money.
提案手法 A	I’m just waiting for the money.
提案手法 B	I’m just waiting for the money.
VMT	Just got us there, but I’m just waiting for the money.
原言語文	助けを求めに来たの
リファレンス	He came to ask for your help.
mBART	I’m here to help you.
提案手法 A	I’m here for your help.
提案手法 B	I came to ask for your help
VMT	He came to me for help.

謝辞

本研究は JSPS 科研費 JP21H05054 の助成を受けたものです

参考文献

[1] Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. Learning to translate in real-time with neural machine translation. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, pp. 1053–1062, Valencia, Spain, April 2017. Association for Computational Linguistics.

[2] Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. Towards fully automated manga translation. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 35, pp. 12998–13008, 2021.

[3] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In **Proceedings of the 5th Workshop on Vision and Language**, pp. 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics.

[4] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In **Proceedings of the IEEE/CVF International Conference on Computer Vision**, pp. 4581–4591, 2019.

[5] Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. Cross-lingual Visual Pre-training for Multimodal Machine Translation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers**, online, April 2021. Association for Computational Linguistics.

[6] Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. Video-guided machine translation with spatial hierarchical attention network. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop**, pp. 87–92, Online, August 2021. Association for Computational Linguistics.

[7] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. 2020.

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **J. Mach. Learn. Res.**, Vol. 21, No. 140, pp. 1–67, 2020.

[9] Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. VISA: An ambiguous subtitles dataset for visual scene-aware machine translation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 6735–6743, Marseille, France, June 2022. European Language Resources Association.

[10] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In **Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)**. NeurIPS, 2018.

[11] Zhishen Yang, Tosho Hirasawa, Mamoru Komachi, and Naoaki Okazaki. Why videos do not guide translations in video-guided machine translation? an empirical evaluation of video-guided machine translation dataset. **Journal of Information Processing**, Vol. 30, pp. 388–396, 2022.

[12] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**, pp. 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[13] Tosho Hirasawa, Zhishen Yang, Mamoru Komachi, and Naoaki Okazaki. Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. **arXiv preprint arXiv:2006.12799**, 2020.

[14] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. 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. 3509–3514, Hong Kong, China, November 2019. Association for Computational Linguistics.

[15] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**, pp. 17949–17958, 2022.

[16] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In **Proceedings of NAACL-HLT 2019: Demonstrations**, 2019.

[17] 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, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[18] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**, pp. 65–72, 2005.

[19] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In **Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing**, pp. 944–952, Cambridge, MA, October 2010. Association for Computational Linguistics.

[20] Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In **Proceedings of the Ninth Workshop on Statistical Machine Translation**, pp. 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.