

令和 7 年度研究開発成果概要書

採 択 番 号 22501
研究開発課題名 自動翻訳の精度向上のための「マルチモーダル情報の外部制御可能なモデリング」
 の研究開発
副 題 マルチモーダル情報理解と制御可能なテキスト生成の研究開発

(1) 研究開発の目的

人間同士および人間とコンピュータ間の円滑なコミュニケーションの実現に向けて、コンピュータが文脈・状況を考慮しながら、言語、音声、画像、動画、表などによるマルチモーダルな情報を統合的に理解する手法を探求する。また、長さや焦点、スタイル、難易度などを指定したり、外部知識を与えることによって、コンピュータが人間のニーズに合わせてテキストを生成する技術を確立する。さらに、これらの成果を統合することで、自然言語生成の新たな応用を開拓するとともに、自動翻訳や同時通訳の研究開発の高度化につなげる。

全ての研究・開発において、深層学習を共通基盤として採用し、タスクやモーダルの垣根を超えた人工知能の実現を目指す。また、自然言語処理、画像処理、深層学習、報道などの分野で最先端の取り組みを進めているグループでチームを結成し、人材やデータ、技術の交流を促進する。研究と並行して言語資源の開発に注力し、その成果物を研究コミュニティに還元する。これにより、日本語の自然言語処理に関する研究で世界トップレベルを維持するとともに、マルチモーダル情報理解や制御可能な自然言語生成などの研究分野で、世界に先駆けた研究を展開する。

(2) 研究開発期間

令和 3 年度から令和 7 年度 (5 年間)

(3) 受託者

国立大学法人東京科学大学<代表研究者>
国立大学法人東京大学
国立大学法人愛媛大学
国立大学法人一橋大学
日本放送協会
株式会社時事通信社

(4) 研究開発予算 (契約額)

令和 3 年度から令和 7 年度までの総額 430 百万円 (令和 7 年度 100 百万円)
※百万円未満切り上げ

(5) 研究開発項目と担当

研究開発項目 1 マルチモーダル情報理解技術の研究開発

- | | | |
|-----|-------------------------|--------------|
| 1-1 | マルチモーダル動画対訳コーパスに関する研究開発 | (国立大学法人東京大学) |
| 1-2 | マルチモーダル機械翻訳に関する研究開発 | (国立大学法人愛媛大学) |
| 1-3 | マルチモーダル情報理解に関する研究開発 | (国立大学法人東京大学) |

研究開発項目 2 制御可能なテキスト生成技術の研究開発

- | | | |
|-----|-----------------|----------------|
| 2-1 | 自動要約の制御に関する研究開発 | (国立大学法人東京科学大学) |
| 2-2 | 翻訳の制御に関する研究開発 | (日本放送協会) |
| 2-3 | スタイルの制御に関する研究開発 | (国立大学法人一橋大学) |
| 2-4 | データ整備に関する研究開発 | (株式会社時事通信社) |

(6) 特許出願、外部発表等

		累計（件）	当該年度（件）
特許出願	国内出願	0	0
	外国出願	0	0
外部発表等	研究論文	13	3
	その他研究発表	154	39
	標準化提案・採択	0	0
	プレスリリース・報道	4	1
	展示会	1	1
	受賞・表彰	19	6

(7) 具体的な実施内容と最終成果

研究開発項目 1：マルチモーダル情報理解技術の研究開発

1-1 マルチモーダル動画対訳コーパスに関する研究開発

映像データを活用するマルチモーダル機械翻訳の研究基盤となるコーパスの研究開発を進めた。翻訳においてマルチモーダル情報が有用と考えられるシーンとして、プレゼンテーションとインタラクションの2つの場면을題材とし、それぞれ映像データとその内容に関連する対訳コーパスからなる動画対訳コーパスを開発した。本研究において開発したコーパスは速やかに公開する必要があるため、権利関係やプライバシー面で懸念のあるデータを用いることは適切ではない。このため、コーパスの作成にあたっては、これらの問題がクリアされている英語圏における既存の動画データセットの英語テキストを日本語翻訳することを基本的な方針とした。マルチモーダル動画対訳コーパスの要件を整理・調査した結果、プレゼンテーション動画対訳コーパスは YouCook2、インタラクション動画対訳コーパスは Audio-Visual Scene Aware Dialog (AVSD) と呼ばれるデータセットをベースとして選定した。英語テキストを日本語に翻訳するときは、必ず動画像情報を参照しながら適切な訳文を導出することとし、動画にグラウンディングされた対訳コーパスになるように留意した。本研究開発項目の成果を研究コミュニティに還元するため、プレゼンテーション動画対訳コーパス YouCook2-JP (75 時間、13,000 文対) を構築・公開し、インタラクション動画対訳コーパス AVSD-JP (8.7 時間、19,000 文対) を構築・公開した。また、これらを用いた国際技術コンペティションとして、WAT 2025 English-Japanese Video Caption Translation Shared Task を実施した。

1-2 マルチモーダル機械翻訳に関する研究開発

機械翻訳への応用を想定し、動画像と言語によるマルチモーダル情報理解技術の研究開発を進めた。プレゼンテーション動画対訳コーパス (WSLT 2017 および TED Talks 原ビデオ) を用いて、5 つ組マルチモーダルスピーチ対訳コーパス (音声・画像・音声認識文・英語・日本語) とスピーチ機械翻訳システムの開発を行い、翻訳性能評価を行った。自動書き起こし文 (音声認識文) の自動修正を用いたスピーチ機械翻訳の研究を行い、音声認識の誤りを訂正することにより、翻訳精度が改善されることを示した。また、潜在拡散モデルによる画像生成を用いたマルチモーダル機械翻訳の研究を行い、原画像を画像生成で修正することにより、マルチモーダル機械翻訳の精度が向上することを示した。さらに、研究開発項目 1-1 で開発した料理動画英日対訳コーパス YouCook2-JP から画像を抽出することでマルチモーダル機械翻訳のための評価用データセット (150 文) を作成し、画像の説明を与える思考の連鎖 (Chain-of-Thought) を用いたマルチモーダル機械翻訳手法を開発した。LLM を用いた実験では、思考の連鎖を用いることで評価用データセットに対する翻訳精度が大きく改善されることを確認した。これらのマルチモーダル機械翻訳の応用先の一つとして、漫画を対象としたマルチモーダル機械翻訳の研究を行い、コマを超えた吹き出し間の文脈情報を扱う機械翻訳手法、および著者情報、出版社情報、ジャンル情報などの書誌情報を用いた機械翻訳手法を提案した。この研究成果論文はトップ国際会議である LREC-COLING2024 に採択された。また、機械翻訳のためのサブワード分割の研究は EACL2026SRW に採択され、Best Paper Award を受賞した。

1-3 マルチモーダル情報理解に関する研究開発

マルチモーダル情報の高度な認識理解を行うための基盤技術として、画像・動画・テキストなどの各モダリティから物体・イベントなどの意味情報を取り出し、これをシーングラフなどの汎用的かつ解釈性のよい表現を用いて統合する方法論と手法を開発した。まず、最も単純な問題である画像識別タスクにおいて、学習に基づくデータ拡張や入力量子化により、テスト時の未知ノイズに対し頑健な画像認識手法を複数開発した。続いて、知識グラフなどの表現を基に LLM エージェントの対話を検証・修正する枠組みを開発し、複数エージェントの対話の一貫性・多様性の向上、パーソナライゼーション、対話の評価などの研究成果を発表した。さらに、マルチモーダル情報処理においてシーングラフや知識グラフを活用する手法として、検索拡張型の画像キャプションの研究を進めた。外部知識グラフをトランスフォーマへ接続した予測キャプション手法はコンピュータビジョンの最難関国際会議である CVPR2023 に採択され、マルチモーダル検索拡張によるオープンワールド画像キャプションは CVPR2024 に採択されており、その実用性や先進性が高く評価された。

研究開発項目2：制御可能なテキスト生成技術の研究開発

2-1 自動要約の制御に関する研究開発

同一の入力に対しても異なる文章を生成できるようにするため、自然言語生成モデルの制御に関する研究を進めた。まず、会議議事録からの要約生成として、クエリ推薦付き自動要約というタスクと手法を提案し、NLP2023 スポンサー賞を受賞した。この研究を日本語で展開することを検討し、自動要約の需要があること、著作権がクリアで後々公開できることを踏まえ、課題 197 で構築したビジネスシーン対話コーパスにクエリと要約を付与したコーパスを構築・公開した。研究開発項目 1 の研究成果を統合するため、学術論文の図表のキャプションの自動生成の研究を進めた。また、大規模言語モデル (LLM) のブームに先駆けて、人間と LLM の対話を通して LLM の性能を高めることを目指した研究を進め、その研究をまとめた論文が NLP2023 委員特別賞を受賞し、トップ国際会議である EACL2024 に採択された。この対話による LLM の制御・性能向上に関する研究を、補助と敵対という異なる視点から高度化させた研究を進め、NLP2024 若手奨励賞を受賞し、論文が ECAI2024 に採択された。さらに、LLM が汎用的な人工知能として普及したことを踏まえ、テキストが LLM によって書かれたかどうかを検出する手法 (LLM 文検出器) を敵対的な文脈内学習で制御する研究を進め、トップ国際会議 AAAI2024 に論文が採択された。その他、自然言語生成の自動評価におけるバイアスの緩和の研究はトップ国際会議 ACL2024 に採択され、2025 年度言語処理学会最優秀論文賞を受賞した。直接選好最適化 (DPO) の問題点とその解決法の提案した研究は NLP2026 最優秀賞を受賞した。このように、自然言語生成の制御に関して、理論・応用の両面において顕著な成果をあげた。

2-2 翻訳の制御に関する研究開発

機械翻訳の品質が一定水準に達し、観点やスタイルなどに応じて翻訳出力を制御する技術への関心が高まっている。本項目では、直前のニュースや知識ベースなどの外部情報を活用し、日英機械翻訳の出力を制御する手法の研究開発を進めた。まず、外部情報を利用した機械翻訳の課題を洗い出すため、日本語と英語のニュース間の情報の差分を分析した。そして、外部情報を利用した翻訳のためのテストデータを構築するため、日本語と英語のニュース間の情報の差分を考慮したテストセット (日英ニュース記事対) を構築し、国際ワークショップ WAT 2023 と国際会議 WMT 2024 で共通タスクを実施した。これらの言語資源を礎として、漸進的アプローチで編集量を制御するニューラル機械翻訳器を提案し、適切な制御パラメータを与えた条件下において、従来手法と比較して有意水準 5% で有意に精度が向上することを確認した。また、文と文のつながりや文書のスタイルを考慮して機械翻訳結果を制御する手法の研究を進め、記事単位での日英ニュース機械翻訳のデータセットを構築し、国際ワークショップ WAT 2025 の共通タスクを実施した。さらに、翻訳出力制御に関する研究開発として、LLM が持っている外部知識を活用した翻訳モデルの研究を進めた。複数の小型 LLM を活用した機械翻訳のための

協力デコーディング（NLP 2025、第 2 回 AAMT 若手翻訳研究会優秀賞受賞）と、低リソース言語における LLM を用いたニュース翻訳のための合成データ生成（NLP 2025、APSIPA 2025）の研究を行い、適切なベースライン手法と比較して生成の精度が有意水準 5%で優れていることを示した。

2-3 スタイルの制御に関する研究開発

文章や発話におけるスタイルの違いを体系的に捉え、それを意図的かつ安定的に制御するための理論と技術の確立を目的とし、文体や難易度などのスタイル要因を定義・モデル化し、文脈や利用目的に応じて適切なスタイルを生成・変換できる手法を探索した。スタイルとしてテキスト難易度を採用し、日本語の文書単位のテキスト平易化のデータセットを構築し、国際コンペを実施した。また、日本語の Wikipedia の文書単位のテキスト平易化コーパスに対して、テキスト平易化システムの出力を生成し、人手で平易度のアノテーションを行うことで、LLM-as-a-Judge を含む自動平易化システムの評価を進めた。さらに、英語学習者向けに英語のテキストの文書単位でのリーダビリティを制御するために、LLM を用いて難易度の制御を行う手法を提案し、国際コンペティションで参加 20 チーム中 3 位の成績を収め、本研究成果の優位性を示した。

2-4 データ整備に関する研究開発

記者・カメラマン等が取材、配信した記事、写真データ、書誌情報などを、単言語、複数言語の対訳・紐づけセットなどの形で整理・蓄積する体制を整え、各機関に定期的に提供した。また、配慮すべきデータ（要配慮個人情報）の取り扱いについて、人間の知見を活用する方法を検討し、有効性を確認した。さらに、日英翻訳の業務フローに機械翻訳と LLM 自動校正を統合し、業務を効率化できることを示した。

(8) 研究開発成果の展開・普及等に向けた計画・展望

本研究開発で得られた成果は、マルチモーダル情報理解および制御可能なテキスト生成という基盤技術として、学術的・産業的に高い波及効果がある。今後は、以下の方針に基づき、成果の展開および普及を進める。

- 本研究の成果物、例えば動画対訳コーパス、テキスト生成コーパス、評価データセット、ソフトウェアについては、メンテナンスと拡張を行い、研究成果のオープン化とコミュニティへの還元を継続する。
- 研究成果の概要を分かりやすく紹介する Web サイト・デモシステムを公開する。
- 国際技術コンペティション、ワークショップ、チュートリアルを開催を継続する。
- 公開講演、一般向けイベント、プレスリリース等を通じた社会への情報発信を継続する。

放送・報道、教育、観光、国際会議運営、行政、企業のグローバル業務支援などの分野において、マルチモーダル翻訳・要約・スタイル制御技術が実運用に組み込まれ、動画・音声・文書を対象とした多言語情報処理の自動化が進捗すると見込まれる。本研究成果により、動画・音声・文書を横断した高品質な多言語情報アクセスが可能となり、国境・言語の壁を越えた情報共有が促進されると期待される。特に、利用者のニーズに応じて内容・長さ・表現を調整できる生成技術は、従来困難であった柔軟な情報提供を実現し、社会全体の情報利便性を大きく向上させる。

公開されたデータセット、ソフトウェア、評価枠組みは、多くの国内外研究者に利用され、マルチモーダル機械翻訳、マルチモーダル LLM、制御可能生成、対話エージェント研究など、関連分野の研究を加速すると期待される。特に、視覚・音声・言語・知識を統合する研究の発展に寄与し、新たな研究領域の形成につながることを期待される。本研究成果は、マルチモーダル LLM、エージェント型 AI、AI による知的支援、AI for Science、教育支援 AI など、次世代研究開発への基盤として活用できる。特に、外部知識や環境情報を統合し、自己修正・自己制御を行う生成モデル研究への展開が期待される。

研究成果の公開と活用を通じて、国内外の研究者・企業・公共機関の連携が促進され、オープンサイエンスの推進、研究データ共有文化の定着、国際共同研究の活性化といった副次的効果も期待される。日本語を中心とした高品質マルチモーダル研究資源の整備は、日本発 AI 技術の国際競争力向上にも寄与する。