『リグ・ヴェーダ』の訳註を用いたヴェーダ語機械翻訳の改善

塚越柚季 [1]　大向一輝 [1]
[1] 東京大学大学院人文社会系研究科
{yuzuki,i2k}@l.u-tokyo.ac.jp

概要

本研究は，翻訳に付随する註釈を言語モデルの学習に統合することで，古代言語翻訳の性能を向上させる手法を提案する．対象として『リグ・ヴェーダ』を用い，現代の学術的な英訳および訳註を活用した．複数の大規模言語モデルに対して教師ありファインチューニングを実施し，標準的な対訳の学習と，原文に訳註を統合した学習とを比較した．BLEU および COMET による評価の結果，比較的大規模なモデルにおいて，訳註を統合することで翻訳性能の向上が確認された．特に，文化的背景を説明する訳註や複雑な文法を分析する訳註が、性能改善に大きく寄与していることが明らかになった．

1　はじめに

サンスクリット語は，古代インドの言語であり，数多くの文献に残る言語である．サンスクリット語最古層のヴェーダ語は，ヴェーダ文献群に残る．古風な語彙，弁別的アクセントの体系，儀礼や文化に関する高度な文脈を伴っており，正確な解釈には言語学的理解にとどまらず，文化的・儀礼的知識を含む深い洞察が求められる．

自然言語処理技術の進展により，多くの言語対において機械翻訳の性能向上が示されてきた．特に，形態的に豊かな言語を対象として，言語学的アノテーションを翻訳モデルに組み込むことで翻訳品質を向上させる研究も行われている．しかし，手動であれ自動であれアノテーション付与が困難であるような古代言語においては，こうした手法の適用は依然として限られている．

本論文では，学術的な訳註を活用して，ヴェーダ語の機械翻訳精度を向上させる手法を提案する．この手法は，古典テキストの翻訳には単なる言語能力だけでなく，歴史的・文化的・儀礼的文脈に対する深い理解が不可欠であるという認識に基づく．こうした知識は，伝統的に数世紀にわたる学術的註釈の中に保存されてきたものである．

2　ヴェーダ文献と訳註

ヴェーダ文献は，およそ紀元前 1500 年から紀元前 500 年に成立した古代インドの宗教文献である．その中でも最古の文献『リグ・ヴェーダ』（Ṛgveda, 以下ṚV）は，種々の神格に捧げられる讃歌の集成である．ṚV に収められた讃歌は韻文から成り，韻律構造による制約や，古風な語彙，ṚV に特有の文法が，翻訳過程を複雑なものにしている．

ヴェーダ文献の翻訳は 19 世紀から出版されてきたが，その中には重要な文献学的・言語学的洞察を与える註釈を伴うものもある．翻訳に付随するこれらの註釈は，ヴェーダの複雑な意味を理解する際に頻繁に参照される．例えば，『リグ・ヴェーダ』だけでも，複数の翻訳が存在する [1, 2, 3, 4, 5, 6, 7, 8]. [1)] このような学術的翻訳（訳文とそれに付随する訳註）自体が，文献学および言語学における重要な成果である．ここで，訳者の解釈とその根拠が提示されている．

例えば，ṚV 8.5.22 [2)] は，[5] で次のように翻訳され，訳註が付与されている．

翻訳：

When did the son of Tugra, abandoned in the sea, do reverence to you, o men, so that your chariot would fly with its birds?

訳註：

The subjunctive pátāt seems to be used in an unusual past prospective sense in this mythological context. This may be an English problem, however. Since the verb of the main clause is injunc. vidhat, this context is not necessarily preterital, but ‘‘timeless,’’ and the subjunctive can therefore be expressing pure future

1)　ここに挙げる参考文献は，『リグ・ヴェーダ』の種々の翻訳を網羅しているわけではなく，代表的なものを示しているに過ぎない．

2)　原文は以下の通りである：*kadā vā taugriyó vidhat samudré jahitó narā yád vā rátho víbhiṣ pátāt*.

modality. The fact that the next verse is also mythological and contains an undoubted present tense form daśasyathaḥ shows that mythological tense is fluid here. Re remarks (ad vs. 23) that the indifference between present and preterite underlines the reflection of the current human situation in the legendary material.

3 手法

3.1 データセット

一般的な対訳コーパスは原文と参照訳から成るが，本研究では，さらに訳註を加えた 3 要素 1 組のデータセットを構築した．参照訳には，現代の英訳 [5] を，訳註には，同著者らによる詳細な訳註 [9] [3] を採用した．翻訳は，『リグ・ヴェーダ』全 1,028 讃歌，1 万を超える全詩節を網羅している．一方，一部の讃歌には，詩節ごとの訳註が存在せず，讃歌全体に対する訳註のみが付されている場合がある．讃歌レベルの訳註も全体解釈にとって有用な文脈を提供するが，本研究では詩節レベルの訳註のみに限定してデータセットを構築した．これは，言語モデルに入力する文脈長を制御し，各学習サンプルを現実的なトークン数の範囲内に収める必要性に基づくものである．

原文には，TITUS（Thesaurus Indogermanischer Text- und Sprachmaterialien）[4] から公開されている電子テキスト [10] を使用した．これは Aufrecht による校訂版 [11] に基づいている．この原文を，IAST 翻字規格 [12] に従ってラテン文字に変換した．現代のサンスクリット研究では，サンスクリットラテン文字表記方式として，IAST と ISO 15919 が広く使用されている．ここで IAST 方式を採用した理由は，英訳および訳註においてサンスクリット語句が IAST で翻字されており，それらとの整合性を確保するためである．

抽出した訳註には，正確な翻訳に不可欠な複数層の文献学的情報が含まれている．具体的には，文法的分析や語彙説明を含む言語学的説明，文化的・宗教的説明，テクスト間関係を扱うテクスト構造の註釈である．

このような多層的内容を含む訳註は，言語学的特徴のみに依存する従来研究とは一線を画す，豊富な文脈情報を提供する．学術的翻訳に付随する註釈は，100 年以上にわたる専門的な文献学的分析の成果であり，アノテーション手法では得られない洞察を多分に含んでいる．

最終的なデータセットは合計 6,754 サンプルから構成されている．これを学習用（5,282 サンプル，78.2%），検証用（743 サンプル，11.0%），テスト用（729 サンプル，10.8%）に分割した．原文は韻文であるため，比較的長さが均一である．訳文は，原文よりも一貫して長い．これは，訳文のなかでも補足説明が必要となるためである．一方，訳註の長さは著しく大きくばらつく．これは，特に複雑あるいは重要な詩節に対しては詳細な分析が与えられる一方，比較的単純な箇所には簡潔な訳註のみが付されるからである．

訳文と訳註との語彙的重複度を定量化するため，n-gram Jaccard 係数（n = 1, 2, 3, 4）を算出した．その平均値は，それぞれ n = 1 で 0.082，n = 2 で 0.018，n = 3 で 0.0075，n = 4 で 0.0038 であった．図 1 は，n-gram Jaccard 係数の箱ひげ図を示している．これらの低い重複値は，訳註から参照訳へのリークのリスクが極めて小さいことを示しており，訳註が訳文の内容を示すものではなく，主として補完的な情報を提供していることを示す．

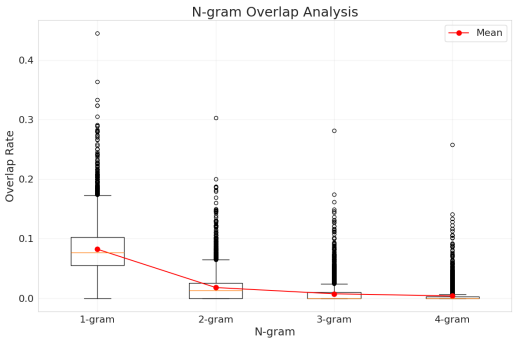

図 1 参照翻訳文と訳註文の間の n-gram Jaccard 係数の箱ひげ図.

さらに，註釈内容の性質を把握するため，1,000 件の訳註を無作為に抽出し，言語学的内容，文化的内容，宗教的内容，哲学的内容，その他の 5 カテゴリから 1 つ以上のラベルを付与した．なお，1 つの訳註に複数ラベルを付与することは許可した．その結果，言語学的内容が 73.4%と最も多く，次いで宗教的内容（32.6%），哲学的内容（15.0%），文化的内容（12.6%），その他（4.0%）が続いた．個々の訳註単位では，54.4%が単一ラベルであった一方，45.6%が複数ラベルを付与されており，訳註のほぼ半数が

3） http://rigvedacommentary.alc.ucla.edu/.
4） https://titus.uni-frankfurt.de.

複数の側面にまたがる統合的な情報を含んでいることが示された．

3.2 モデル選択

本研究では，異なるアーキテクチャ，パラメータ規模，学習パラダイムにわたる汎化性能を評価するため，以下の 5 種類の多様な大規模言語モデルを用いて提案手法を検証する．**GPT-4.1 nano**[5]，**Gemini 2.5 flash**[6]，**Gemma 2 Mitra**[7]，**Llama 3.1 8B**[8]，および **Llama 3.2 3B**[9] である．Gemma 2 Mitra はサンスクリット語に特化したモデルである．

3.3 学習手法

本研究では，翻訳性能向上の課題を条件付きテキスト生成とみなし，以下の学習条件を比較する．

標準 SFT：ヴェーダ文献テキストを入力として，英語翻訳を生成するようモデルをファインチューニングする．

註釈拡張 SFT：ヴェーダ文献テキストと英語による学術的註釈の両方を入力として，英語翻訳を生成するようモデルをファインチューニングする．

註釈のみ SFT：註釈テキストのみを入力として，英語翻訳を生成するようモデルをファインチューニングする．

プロプライエタリモデル（GPT-4.1 nano，Gemini 2.5 flash）およびオープンウェイトモデル（Llama，Mitra）の双方について，それぞれのプラットフォームに適した最適化手法を用いてファインチューニングを実施した．詳細な学習設定，ハイパーパラメータ，ならびに入力フォーマットの仕様については，付録 A に記す．

3.4 評価指標

翻訳品質の評価には，BLEU および COMET スコアを用いた．BLEU スコアは，標準化され再現性のある評価を行うため，SacreBLEU [13] を用いて算出した．COMET スコアは，事前学習済みの wmt22-comet-da モデルを用いて算出した [14]．

5) https://platform.openai.com/docs/models/gpt-4.1-nano
6) https://ai.google.dev/gemini-api/docs/models#gemini-2.5-flash
7) https://huggingface.co/buddhist-nlp/gemma-2-mitra-it
8) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
9) https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

表 1 モデルおよびデータセット構成ごとの BLEU および COMET のスコア

学習データ	テストデータ	BLEU-4	COMET
GPT-4.1 nano			
訳+註釈	訳+註釈	4.2	0.61
訳	訳	1.6	0.58
訳	訳+註釈	1.7	0.59
註釈	訳	0.39	0.52
Gemini 2.5 flash			
訳+註釈	訳+註釈	14.4	0.60
訳	訳	11.3	0.67
訳	訳+註釈	13.6	0.69
註釈	訳	5.1	0.62
Gemma 2 Mitra			
訳+註釈	訳+註釈	6.5	0.63
訳	訳	4.9	0.59
訳	訳+註釈	2.3	0.52
註釈	訳	1.8	0.56
Llama 3.1 8B			
訳+註釈	訳+註釈	7.3	0.62
訳	訳	5.6	0.63
訳	訳+註釈	5.4	0.61
註釈	訳	1.2	0.56
Llama 3.2 3B			
訳+註釈	訳+註釈	3.5	0.59
訳	訳	3.9	0.60
訳	訳+註釈	2.8	0.57
註釈	訳	0.27	0.52

4 結果

4.1 結果：BLEU および COMET スコア

表 2 註釈拡張 SFT による BLEU スコアの向上率

モデル	向上率 (%)
GPT-4.1 nano	+155.0
Gemini 2.5 flash	+27.4
Gemma 2 Mitra	+30.5
Llama 3.1 8B	+31.1
Llama 3.2 3B	-11.5

表 1 に，3 つの実験設定の下で評価した各モデルの BLEU および COMET スコアを，表 2 に，対訳のみによる学習を基準にした，訳註を統合した学習の BLEU スコアの向上率を示す．ほとんどの設定において，訳註を統合することで BLEU および COMET

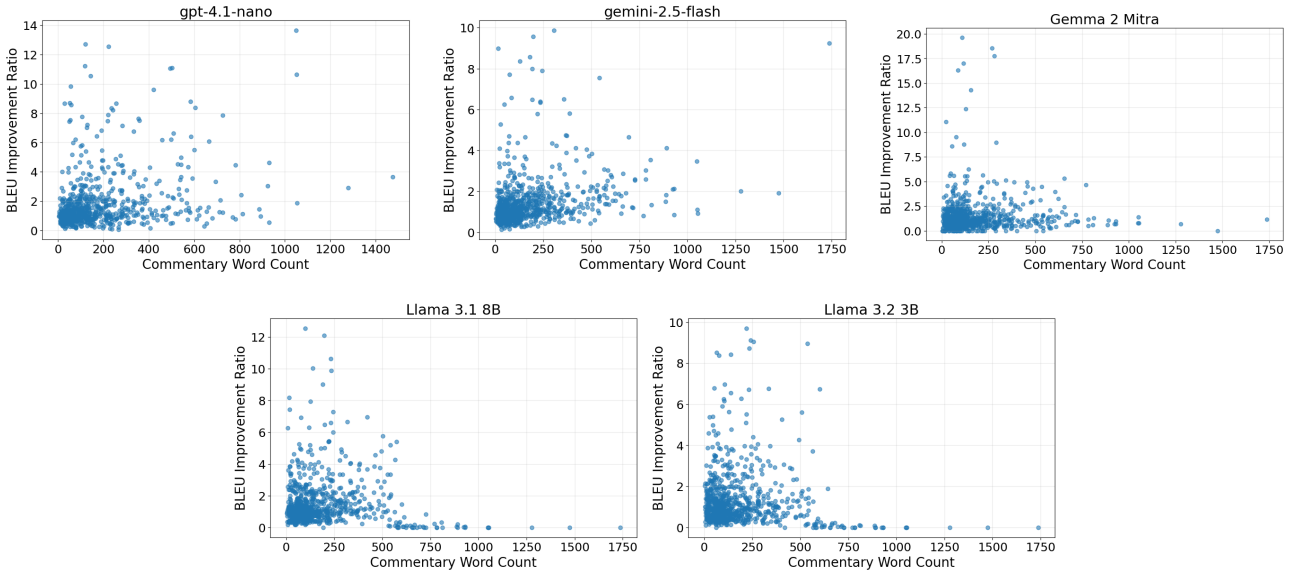

図 2　モデル別の BLEU 向上率と註釈長との関係を表す散布図．各点は一つの詩節を表す．

スコアの双方が向上しており，語彙的正確性および意味的整合性の改善が示されている．しかし，Llama 3.2 3B では BLEU および COMET の双方においてわずかな性能低下が見られ，小規模モデルにおいては訳註の付加が限定的な効果しかもたらさない可能性が示唆される．

大規模なオープンソースモデルである Gemma 2 Mitra および Llama 3.1 8B も，それぞれ +30.5%，+31.1% と大きな性能向上を示している．一方，Llama 3.2 3B では，註釈を統合した場合に BLEU-4 スコアが -11.5% 低下しており，前述の結果と同様の傾向が確認された．

さらに，訳註の長さが翻訳性能の改善と関連しているかも検討した．図 2 に示すように，短い訳註と長い訳註のいずれにおいても性能向上が見られる場合がある．このことから，訳註の長さそのものが，観測された性能改善の主たる要因ではないことが示唆される．

4.2　定性的分析

ヴェーダ語は高度に屈折的であり，形態論的・意味論的曖昧性が翻訳における主要な困難となる．訳註は，音韻変化，語幹同定，活用パターン，語義の限定といった言語学的情報を明示的に提供し，モデルが最も一般的あるいは字義的な解釈に偏ることなく，文脈に即した訳を生成することを可能にしている．さらに，訳註は，宗教的，神話的背景といった文化的前提知識を補完する点で重要である．こうした情報はヴェーダ文献の文化的背景を明示化し，単なる言語内容を超えた適切な翻訳生成に寄与している．

性能向上が確認されたモデル群において BLEU スコアが一貫して向上したことは，註釈統合が語レベルの正確性にとどまらず，フレーズおよび文レベルの一貫性を改善していることを示している．この傾向は，訳註が提供する構造的・文脈的情報の性質と整合的である．一方で，訳註の語数と BLEU 改善との間に明確な関係はなく，性能向上の決定的要因は訳註の長さではなく訳註の質であると考えられる．特に，明示的な文法分析，曖昧性解消の手がかり，文化的背景説明といった要素が有効である．

5　結論

本研究は，学術的な翻訳に付随する註釈を取り入れることで，複数のモデルにわたってヴェーダ語翻訳の精度が向上することを示した．本研究で提示した註釈統合の枠組みは，古代言語の機械翻訳において学術的専門知識を活用する一つの方法として位置づけられる．また，古代言語に限らず，特定ドメインの知識が重要となる翻訳タスクに対しても応用の余地がある．今後の課題として，註釈そのものを自動的に生成・選別する手法の検討や，註釈利用の効果を適切に測定する評価軸の確立が挙げられる．また，註釈統合の有用性が，他の文献や同一文献に対する異なる翻訳・訳註資源においても再現されるかについては，さらなる検証が必要である．

謝辞

本研究は JSPS 科研費 JP25K21518 の助成を受けたものです.

参考文献

[1] Hermann Grassmann. **Rig-Veda : übersetzt und mit kritischen und erläuternden anmerkungen versehen von Hermann Grassmann**. Brockhaus, Leipzig, 1876.

[2] Karl F. Geldner. **Der Rig-Veda : aus dem Sanskrit ins Deutsche übersetzt und mit einem laufenden Kommentar versehen**. No. v. 33-36 in Harvard oriental series. Harvard University Press , Oxford University Press , Otto Harrassowitz, Cambridge, Mass. , London , Leipzig, 1951.

[3] Louis Renou. **Études védiques et pāṇinéennes**. No. sér. in-8o ; fasc. 1-2, 4, 6, 9-10, 12, 14, 16-18, 20, 22-23, 26-27, 30 in Publications de l'Institut de civilisation indienne. E. de Boccard, Paris, 1955.

[4] Tat'jana Jakovrevna Elizarenkova. **Rigveda**. 第 3 巻 in Literaturnye pamjatniki. Nauka, 1999.

[5] Stephanie W. Jamison and Joel P. Brereton. **The Rigveda: The earliest religious poetry of India**. Oxford University Press, New York, 2014.

[6] Michael Witzel, Toshifumi Gotō, Eijirō Dōyama, and Mislav Ježić. **Rig-Veda : das heilige Wissen Erster und zweiter Liederkreis**. Verlag der Weltreligionen, Frankfurt am Main, 2007.

[7] Michael Witzel, Toshifumi Gotō, and Salvatore Scarlata. **Rig-Veda : das heilige Wissen Dritter bis fünfter Liederkreis**. Verlag der Weltreligionen, Frankfurt am Main, 2013.

[8] Eijirō Dōyama and Toshifumi Gotō. **Rig-Veda: das heilige Wissen: sechster und siebter Liederkreis**. Verl. der Weltreligionen, Berlin, 1. aufl edition, 2022.

[9] Stephanie W. Jamison and Joel P. Brereton. Rigveda translation: Commentary, 2015. Center for Digital Humanities, University of California, Los Angeles.

[10] Francisco Javier Martínez García and Jost Gippert. Thesaurus indogermanischer text- und sprachmaterialien, 1995.

[11] Theodor Aufrecht. **Die Hymnen des Rigveda**. Bonn: Adolph Marcus, 1877.

[12] Royal Asiatic Society of Great Britain and Ireland. **Transliteration report**. London : The Society, 1896.

[13] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics.

[14] Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, **Proceedings of the Seventh Conference on Machine Translation (WMT)**, pp. 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In **International Conference on Learning Representations**, 2022.

A　学習設定の詳細

本節では，実験に用いた入力形式および主要な学習設定のみを簡潔にまとめる．

A.1　入力プロンプト形式

註釈拡張 SFT では，原文・註釈・訳文を以下の順で連結した入力を用いた．

```
# Input:
[Original Vedic Sanskrit text]
# Commentary:
[Scholarly philological commentary]
# Translation:
[Target English translation]
```

標準 SFT では註釈を含めず，原文と訳文のみを用いた．推論時にはいずれの場合も `[Target English translation]` 行は含まれない．

A.2　ハイパーパラメータ

プロプライエタリモデル（GPT-4.1 nano, Gemini 2.5 flash）については，各提供元の公式 Fine-tuning API を用い，少数エポックによる軽量な微調整を行った．オープンウェイトモデル（Llama, Gemma）では計算効率のため LoRA [15] を採用し，$r = 16$ の低ランク適応を全アテンション・FFN 層に適用した．学習率は 3×10^{-4}，AdamW オプティマイザを用い，早期終了を有効化した．

B　翻訳が改善した例

本節では，註釈の統合によって BLEU スコアが最も大きく改善したテスト例を示す．

改善例 1（GPT-4.1 nano）

BLEU 向上率: 51.03 (2.63 → 53.66)

詩節: RV 6.2.7

参照訳

For now you are to be reverently invoked among the clans, as our dear guest, to be protected like a son who brings delight to the home, (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

註釈なし

For here you are, o widely spread (flame), the fire to be kindled, dear guest to us, delighting like an old man in his fortress, like a son who is to be protected.

註釈あり

For you are eager for battle, dear guest; you are to be protected like a son who brings delight (to the home), (but also) to be enflamed like a battle-lusty (warrior) in a stronghold.

改善例 2（GPT-4.1 nano）

BLEU 向上率: 33.0282 (1.7544 → 34.7826)

詩節: RV 9.73.1

参照訳

In the jaw (of the pressing stones) while the droplet was blowing its blast, they sounded in unison. The ties of lineage have joined together in the womb of truth. The lord [=Soma] has made his three heads to be seized. The boats of the trusty one [=Soma] have brought the good (ritual) performer across to the other shore.

註釈なし

Streams of soma roar as they cross the filter together, the womb of truth together with the navel of truth, the three heads of the lord of the night (being) joined together. The boat of truth, well made, has been set in motion. The trusty boat has been set in motion.

註釈あり

The streams of soma roar together as they cross the filter, and the ties of lineage have joined together in the womb of truth. The demon made his three heads to be seized. The boat of truth, well made, has been set forth.

改善例 3 (GPT-4.1 nano)

BLEU 向上率: 26.4069 (19.0476 → 45.4545)

詩節: RV 8.75.11

参照訳

Surely you will toil for wealth for us, for our quest for cattle, Agni? Make a wide (way) for us, you wide-maker.

註釈なし

Make for us a dwelling place, Agni, for our quest for cattle, for wealth. Make us strong, strong.

註釈あり

O Agni, you will toil for wealth for us, for our quest for cattle. Make it broad and broad for us.