（報道発表資料）

2026.4.22

NTT 株式会社

LLM 間の「語彙の壁」を克服する世界初の「トークン共通化」技術を確立

～異種 LLM 同士も密に連携させ、高性能化につながる知識の統合や転移を可能に～

発表のポイント：

- ◆ 大規模言語モデル（LLM）の推論における入出力単位「トークン」の語彙集合を、推論中に精度劣化なく自在に縮小できる、世界初の理論およびアルゴリズムを確立しました。
- ◆ 本技術により、任意の異種 LLM 間で共通の語彙集合を介した連携が可能になりました。
- ◆ 本技術をアンサンブルや NTT 独自のポータブルチューニングなどの連携技術に適用することで、より多様な異種 LLM 間で知識の統合・転移が実現できるようになります。

NTT 株式会社（本社：東京都千代田区、代表取締役社長：島田 明、以下「NTT」）は、大規模言語モデル（LLM）における入出力単位「トークン」の語彙を精度劣化なく縮小させ、異なる LLM 間でもトークン語彙を共通化できる世界初の推論技術を確立しました。これまで、複数の LLM を用いてアンサンブル[*1]に代表される推論時連携を実現するには、各 LLM のトークン語彙が一致している必要がありました。本技術によりその制約が解消され、任意の異種 LLM 間で、これまで困難だったアンサンブルや NTT 独自のポータブルチューニング[*2]など様々な推論時連携が可能となり、知識の統合・転移による高精度化を実現できるようになりました。本成果は、2026 年 4 月 23 日から 27 日まで、ブラジル・リオデジャネイロで開催される深層学習分野における最難関国際会議 International Conference on Learning Representations (ICLR) 2026[*3]において発表されます。

1. 背景

近年、大規模言語モデル（LLM）は自然言語で推論を行える AI として急速に活用が進んでいます。LLM は文章を１文字ずつ出力する代わりに、「トークン」と呼ばれる部分単語の単位で効率的に推論を行います。より具体的には、次に出力すべきトークン候補を確率値で予測する「次トークン予測」を都度計算し、この予測結果に基づいたトークンの出力を繰り返すことで推論を進めていきます。このトークン候補の集合は「トークン語彙」と呼ばれ、数万から数十万ものトークンから構成されますが、とくに開発組織や時期の異なる LLM の間では、それらのトークン語彙は一致しないことが一般的です。一方でこのように LLM 間でトークン語彙が異なると、互いに推論中の次トークン予測結果を比較・参照することができません。この「語彙の壁」により、たとえば複数 LLM の予測結果を統合して推論精度を向上させるアンサンブルや、別モデルに専門知識を転移させるポータブルチューニングなど、様々なトークンレベルの連携技術を異種 LLM 間で活用することが困難となっていました。

2. 研究成果の概要

本研究では、LLM が推論に用いるトークン語彙を、精度劣化なく自在に縮小できる世界初の技術を確立しました（図1）。具体的には、LLM の推論中に計算される全トークンに対する次トークン予測結果を、指定された一部のトークン（部分語彙）のみを候補とする予測結果に都度変換していきます。この変換において最終的に出力される文章全体の傾向が変化しないよう、独自の理論に基づいて変換アルゴリズムを設計したことで、元の LLM の推論精度を劣化させることなく、任意の部分語彙での推論が可能になりました。この技術を応用することで、異なる語彙を持つ LLM 間において、それらの「最大共通語彙」上での推論時連携が可能になります（図2）。すなわち、アンサンブルによる知識統合やポータブルチューニングによる知識転移など、従来はトークン語彙の不一致が障壁となっていた推論時連携が、異種 LLM 間でも共通トークンを介して実現できるようになりました。また実験においても、トークン語彙の異なる LLM 同士を対象としたアンサンブルの検証を行い、各 LLM の性能を維持したまま共通トークンによる連携が可能であること、およびその連携により推論精度の向上が実現できることを確認しました。

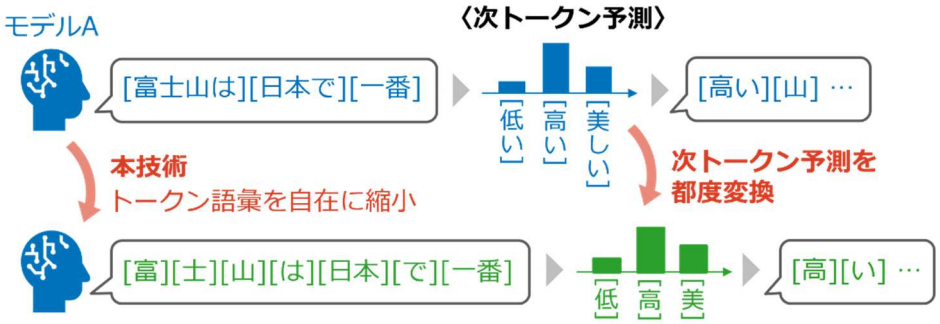

図 1　LLM の出力傾向を保ったまま、
トークン語彙を自在に縮小できる変換技術を確立

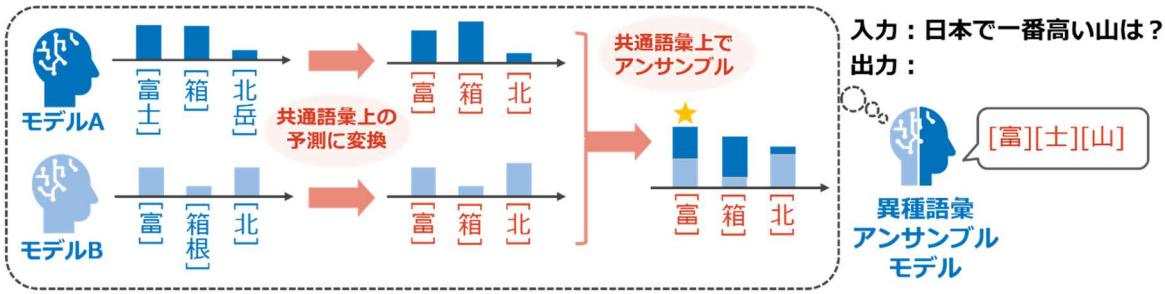

図 2　異なる語彙をもつ LLM 同士でも、共通語彙を介したアンサンブルが可能に

3. 技術のポイント

①「損失なし」で語彙を自在に縮小させる新理論

LLM の出力品質を維持したまま、より少ないトークン候補（部分語彙）での次トークン予測を可能にする、世界初の理論を確立しました。ここで単にトークン候補を減らすだけであれば、削りたいトークン候補の情報を次トークン予測から削ることで実現できますが、本来出力されるはずだった文字

列が出力されなくなることで、大幅な性能劣化につながってしまいます。本研究では、元の語彙に関するトークン確率と部分語彙に関するトークン確率とを統一的に扱える理論的枠組みを構築し、最終的な出力文字列の分布が不変となるために、それらトークン確率の間で満たされるべき関係式を導出しました（図 3）。この関係式に基づいて元の語彙による次トークン予測を変換していくことで、出力品質は変えずに、部分語彙による次トークン予測が計算可能となりました。

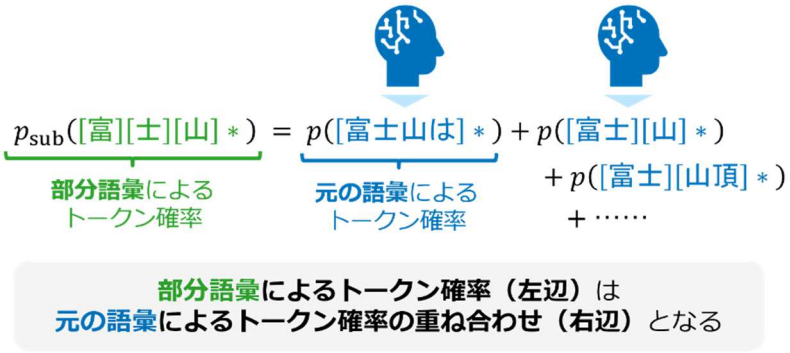

図 3　出力される文章が変化しないために要請される、元の語彙と部分語彙に関するトークン確率の関係式（イメージ）

② 実用的な変換アルゴリズムを導出

ポイント①の理論により任意の部分語彙での次トークン予測が計算可能となりましたが、そのためには元の LLM による複数の次トークン予測結果が必要となり、その計算コストが新たな課題として生じました。この課題に対して本研究では、計算上必要となる複数の次トークン予測のほとんどは過去の計算にも現れるためキャッシュにより再利用できる点、および確率値がほぼ0となる下位トークンは計算上省略できる点を設計に取り入れました（図 4）。これにより、元の語彙による通常の推論時と同程度の計算コストで動作する、効率的な変換アルゴリズムを実現しました。また本アルゴリズムを実際の LLM に適用し、理論通りに出力傾向を保ったまま、様々な部分語彙での推論が可能となることを実験的に確認しました（図 5）。

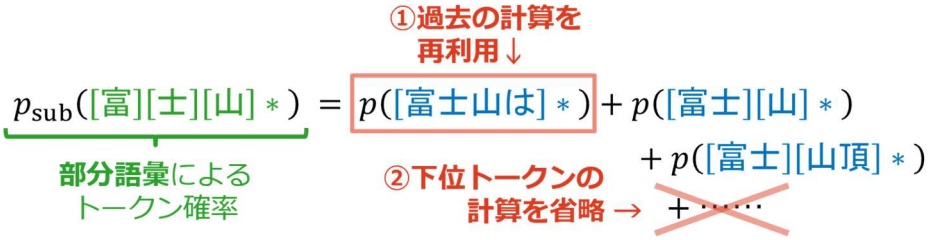

図 4　ポイント①の関係式を効率的に計算するアルゴリズムを実現

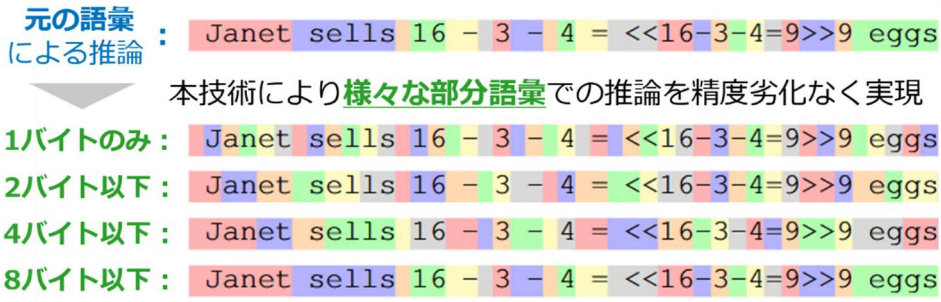

図 5　実際の LLM に本技術を適用し、指定されたバイト数以下のトークンのみを用いて推論させた結果の例

③ 最大共通語彙による異種 LLM 間の知識統合を実現

上記の変換アルゴリズムの応用として、異なる語彙を持つ複数の異種 LLM に対し、それらに共通するトークン候補を全て集めた「最大共通語彙」で推論させることで、異種 LLM 間で次トークン予測の共有が可能となりました。例えばある市中モデル同士の組み合わせにおいては、一方の LLM が約 15 万個、もう一方の LLM が 13 万個のトークンをもち、それらの最大共通語彙は約 6 万ものトークンからなる集合となります。このような共通語彙上で各 LLM を推論させることで、推論精度や速度を犠牲にすることなく、アンサンブルによる知識統合やポータブルチューニングによる知識転移など、次トークン予測の共有を必要とする様々な推論時の連携が異種 LLM 間で実現可能となりました。また実際に語彙の異なる LLM 同士に対して共通語彙によるアンサンブルを適用した実験において、出自の異なる知識の統合により単体モデルに比べて推論能力が大幅に向上することが確認できました（図 6）。

▼単体モデル	GSM8K	MATH
市中モデルA	71.27%	27.86%
市中モデルB	76.65%	26.14%
▼上記モデルのアンサンブル		
理論保証のない既存手法	27.60%	19.57%
本手法（最大共通語彙）	**81.12%**	**30.29%**

図 6　異種 LLM 間の知識統合による数学タスクでの精度向上例

4. 今後の展開

本研究成果により、LLM の推論時の語彙を自在に縮小できるようになり、その応用として異なるトークン語彙をもつ異種 LLM 同士でも次トークン予測レベルでの密な連携が可能となりました。本技術をアンサンブルやポータブルチューニングなど様々な連携技術と組み合わせることで、多様な異種 LLM 間での知識の統合や転移が可能になります。特に NTT の tsuzumi[*4] など独自の語彙をもつ LLM でも、本技術の活用によって他の市中 LLM との連携が容易になっていくことが期待できます。

発表について：

本成果は、2026 年 4 月 23 日～27 日に開催される深層学習分野における最難関国際会議 ICLR 2026（International Conference on Learning Representations）にて、下記のタイトル及び著者で発表されます。

タイトル：“Lossless Vocabulary Reduction for Auto-Regressive Language Models”
著者：千々和 大輝、山口 真弥、大庭 知也、坂尾 珠和、竹内 亨（コンピュータ&データサイエンス研究所）、長谷川 拓、西田 京介（人間情報研究所）
URL：https://openreview.net/forum?id=xAvqHtLVgz

【用語解説】

※1. アンサンブル … 複数のモデルから出力候補の確率値を集約し、モデル間で合意の取れる出力を行う古典的な連携技術。

※2. ポータブルチューニング … 予め学習させておいた報酬モデルを基盤モデルと連携させる方式により、新たな基盤モデルに対しても再学習を行うことなく特化学習の効果を付与できる NTT の独自技術。

・動画解説「生成AIのカスタマイズコストを抜本的に削減するポータブルチューニング技術」https://youtu.be/zQw7ayQknFY

・報道発表（2025 年 7 月 9 日）https://group.ntt/jp/newsrelease/2025/07/09/250709a.html

※3. ICLR2026 … 深層学習に関するトップレベルの国際会議。https://iclr.cc/Conferences/2026

※4. tsuzumi … NTT 版大規模言語モデル「tsuzumi 2」https://www.rd.ntt/research/LLM_tsuzumi.html

■本件に関する報道機関からのお問い合わせ先
NTT 株式会社
サービスイノベーション総合研究所
広報担当
問い合わせフォームへ