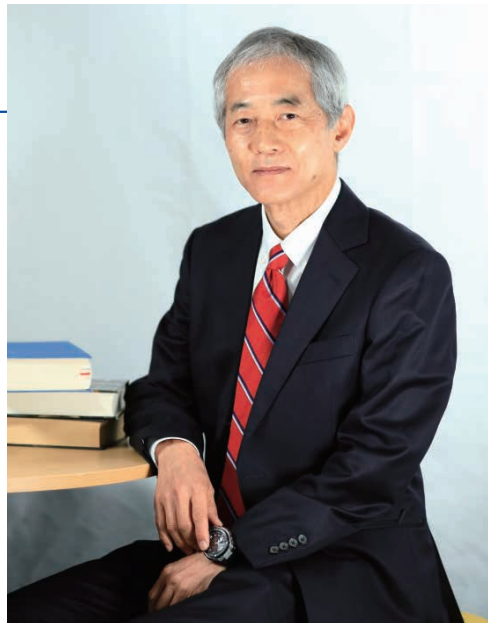
URL https://journal.ntt.co.jp/article/40041
DOI https://doi.org/10.60249/26081101

NTTコミュニケーション科学基礎研究所
上席特別研究員

永田昌明 Masaaki Nagata



LLM時代の機械翻訳の現在地——「次の単語の予測」から「推論」へ

ChatGPTやGeminiなど，いまや大規模言語モデル（LLM）を搭載した生成AIによる翻訳は日常的に使われています．その背景には，ルールベース翻訳から統計翻訳，ニューラル翻訳を経てLLM翻訳へ至る長い開発史があります．その変遷を研究の最前線で見続けてきた永田昌明上席特別研究員に，LLMにおいて何がブレークスルーとなったのか，また「推論モデル」や「AIエージェント」が主流となる中で，どのような研究をしているのか伺いました．

統計翻訳からニューラル翻訳，そしてLLMへ

ご研究の分野について教えてください．

私は統計的機械翻訳の研究を経て，2014年ごろからニューラル機械翻訳が主流になると，日英対訳コーパスや対訳文の単語対応を求めるプログラムの研究に携わってきました．なぜ，こうした研究をしてきたのかをお話しする前に，簡単に機械翻訳の技術史を振り返らせてください．

機械翻訳の研究が始まったのは，古くは1950年ごろのことで，最初は「ルールベース」と呼ばれる手法が使われていました．日本語は，「主語-目的語-動詞」，英語は「主語-動詞-目的語」といったように語順に違いがあります．したがって，目的語と動詞をひっくり返して，単語を辞書で置き換えれば翻訳できる，と考えたわけです．この研究に人類は半世紀も費やしましたが，人間の言語表現は多彩なため人手でルールを網羅するのは困難で，結局，うまくいきませんでした．

次に出てきたのが，IBMのワトソン研究所から始まった「統計的機械翻訳」で，私自身も1990年代から研究を始めました．これは数十万～数百万文の対訳データから，「どの単語がどの単語に対応するか」「どの語順がどれくらい出現するか」を統計的に学習する方法です．2006年からサービスを開始したGoogle翻訳などで比較的うまくいったものの，日本語から英語など，語順が大きく異なる言語では精度が上がりませんでした．

その後，2016年にGoogleが発表した，これまでとは全く異なる「ニューラル翻訳」で精度が格段に上がり，人間に近づいたと世間を驚かせました．これは，エンコーダ・デコーダモデル（符号化器-復号器モデル）といって，入力を処理する「エンコーダ」と，出力データを生成する「デコーダ」に再帰ニューラルネットワーク（RNN：Recurrent Neural Network）を使い，単語を1つずつ入力して内部状態を更新しながら次の単語を予測する手法です．このネットワークを2つ（翻訳元言語と翻訳先言語）接続することで，例えば日本語文を入力したら英語を一単語ずつ予測して出力できるようになりました．

ニューラル翻訳で画期的だったのは，単語の意味情報をベクトル，つまり数字で表現したことにあります．すべての単語の意味を約1000次元という多次元のベクトル空間の1つの点として表しています．そして，たくさんの対訳データを使って訓練することで，「これはペンです」と“This is a pen”や「私は日本人です」と“I am a Japanese”といっ

た文の意味ベクトルが距離的に近い位置に配置されるようになります．これにより，元の文意を汲むような流暢な訳が可能になりました．なお，このニューラル翻訳を開発したイリヤ・スツケヴァー氏は，2024年にノーベル物理学賞を受賞したジェフリー・ヒントン教授の教え子であり，後にChatGPTを生み出したOpenAIの共同設立者です．

RNNを用いたニューラル翻訳は革新的でしたが，並列化が難しく，時間がかかるうえ，長文の翻訳も苦手でした．この問題を解消したのが，2017年にGoogleの研究者らが発表したTransformerです．これは，注意機構（Attention Mechanism）という仕組みを使って，文に含まれる単語どうしの関連性を，同時並列で処理するニューラルネットです．これにより，10億文といった大量のデータを扱えるようになりました．実はこのTransformerのデコーダ（次の単語を予測するネットワーク）だけを取り出して巨大化し，大量のテキストデータで訓練したものが現在の大規模言語モデル（LLM：Large Language Model）[*1]なのです．

ちなみに，このデコーダが巨大になればなるほど，また学習するデータ量や計算量が増えれば増えるほど，性能が向上することを2020年にOpenAIが明らかにしています（スケーリング則）．つまり，LLMの精度向上には莫大なリソースが不可欠なのです．大航海時代に例えるなら，地球は丸いと信じて西回りの航海に臨んだコロンブスがOpenAI，資金を援助したスペインのイサベル女王がマイクロソフトということになるのかもしれません．

さて，こうした機械翻訳の開発史の中で，私自身，時代ごとに重要な研究テーマを見出してきました．ルールベース翻訳から統計翻訳への移行期には「形態素解析」といって，自動で正しく単語を区切るアルゴリズムを開発しました．これは後にかな漢字変換に応用されました．ニューラル翻訳では，対訳データの収集が重要であることから，4000万対を超える日英対訳コーパスを集めたJparaCrawl[*2]や，3億文対を超える特許に関する日英対訳コーパスの作成を主導しました．近年は，研究開発用に小型のLLMが公開されていることから，これらを使ってLLMによる翻訳精度を高める研究をしています．

LLMを使った機械翻訳でいかに精度を高めるか

最近の研究について教えてください．

LLMは，①大量のデータから次単語予測を学ぶ「事前訓練」，②想定問答から正解を学ぶ「指示チューニング」，③適切な出力のための調整「アライメント」という3つの訓練ステップを経て学習します．例えば②では，「日本で梅雨がないのは北海道とどこか」→「小笠原諸島」といったような指示と回答をセットで学習します．また，③では「爆弾のつくり方を教えて」と問われた際に，「爆弾をつくるのは違法なのでお答えできません」といったように，人間社会に適応すべく調整を繰り返し行います．なお，アライメントでは正解を教えるのではなく，LLMの答えの良し悪しを，人間が判断して学習させる「強化学習」を用います（**図1**）．

そして私はここ数年，このLLMの学習の仕組みを機械翻訳に応用する研究を行ってきました．最初は対訳データを使って指示チューニングで精度を上げようと試みたのですが，あるレベルまで到達すると，その後どれだけ対訳データを与えても賢くならないという問題に直面しました．そこで試行錯誤の結果，機械翻訳のシステムを賢くするには，次の単語を予測する事前訓練の段階で，大量の対訳データを与えると良いことを発見したのです（2024年，筑波大との共同研究）．この研究に関しては，ほぼ同時に3つの研

＊1 大規模言語モデル：10億個（one billion:1B）を超える膨大なパラメタを持つニューラルネットを使って大規模なテキスト（>100億トークン,10B）から学習された言語モデル．自然言語の理解や生成に優れています．

＊2 JparaCrawl：世界中で公開されているWebページを大規模に巡回・収集し，そのデータを公開している非営利プロジェクトCommon Crawlの成果を利用して，日本語と英語および日本語と中国語の対訳データを大量に集めるNTTコミュニケーション科学基礎研究所のプロジェクト．

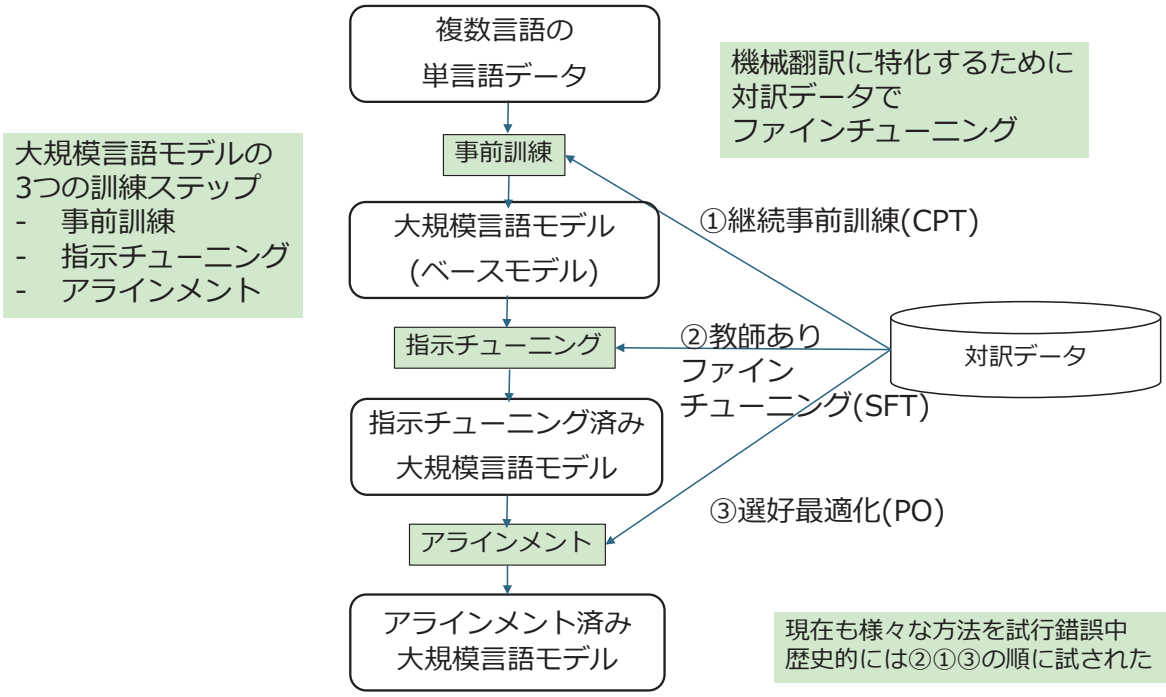


図 1　大規模言語モデルを用いた翻訳（LLM 翻訳）

究グループが同内容の論文を発表したことから，3 カ月の差で別のグループに先を越されてしまいました．論文提出の半年前には気付いていたのですが，検証に時間をかけてしまいました．まさにこの世界はスピード勝負だと痛感しました．

次に取り組んだのが，LLMで問題とされるハルシネーション*3を防ぐためのアラインメントの研究（「単語対応を用いた選好最適化」2024年，東大との共同研究）です．例えば，「There was some damage to cars and buildings in Belgrade due to the snow」という文が，「セルビアの首都ベオグラードは雪で被害を受けた」と訳されたとすると，「セルビア」が勝手に生成されたうえ，「車や建物に」という部分が抜け落ちたことが分かります．そこで，入力文の単語が出力文に訳出された割合である単語カバー率を求める技術を開発しました．これには，以前から私たちが取り組んできた，世界でもっとも精度の高い単語対応の技術が役立ちました．これを使って単語カバー率が高いものを良しとする強化学習を行うことで，「訳抜け」や「湧き出し」を軽減しました．この論文はよく読まれ，今も多く引用されています（**図 2**）．

なお，この分野はさらに進んでいて，今は自動品質評価，すなわち翻訳の良し悪しについてLLMを使って判断する研究が注目されています．私たちも特許の「請求項」*4の翻訳のための自動評価プログラムを開発して評価を行っています．請求項は一文で書かれていて構造も複雑なため翻訳が非常に難しいのですが，権利範囲を明確にするために極めて正確な訳が求められます．このように絶対に間違いが許されない文の翻訳に，LLMによる品質評価や強化学習を適用して，翻訳精度を上げる研究に取り組んでいるところです．

＊3 ハルシネーション：LLMにおいて，原文にない語が湧き出したり，もっともらしい嘘や間違った文をつくり出したりすること．
＊4 請求項：出願人が発明の内容を具体的に特定し，その独占権の範囲を区分けして記載する項目のこと．

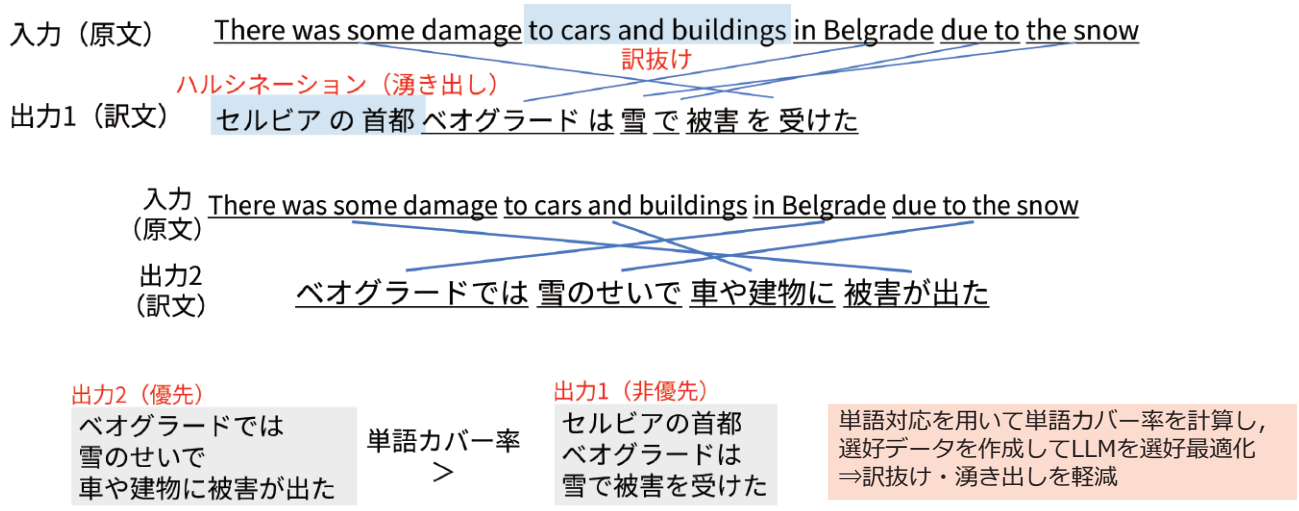


図 2　単語対応を用いた選好最適化

急速に進展する技術を追いながら，諦めることなく新たな研究を

どのようなご苦労があったのでしょうか．

実は2020年にGPT-3が開発された際には，Transformerが巨大化しただけなので，さほど衝撃を受けなかったのです．ところが2024年9月にOpenAIから「o1（オーワン）」というLLMが発表されたときには，世界がガラっと変わったと感じました．これは，推論モデル*5と呼ばれるもので，LLMに思考過程をしゃべらせながら回答させ，試行錯誤を繰り返しながら回答精度を高めるという画期的な手法です．

さらに翌2025年1月には，中国のAI企業DeepSeekから「R1」というLLMが発表されると，中国のベンチャーが米巨大テックよりはるかに低いコストで高性能モデルをつくったとして「DeepSeekショック」が起こりました．さらに，2026年に米・アンソロピックがシステムの脆弱性を探索するAIモデル「Claude Mythos」を発表すると，話題になると同時にソフトウェアのセキュリティを脅かすとして衝撃が走りました．このように，LLMの世界は数カ月ごとに新しい技術が登場し急速に変化していくため，私もその動きを常にキャッチアップしなければなりません．実のところ，私自身は2カ月に一度くらいの頻度で4～8件の論文の査読を頼まれるため，否応なく最新の技術を学ぶことになります．

とはいえ，新技術を試行錯誤しながら理解し，使いこなすころにはまた新しい技術が出てくる．2年後に自分がいったいどんな研究をしているのか，全く想像がつかないのが実情です．

今後の課題と，後進へのアドバイスをお願いします．

昨年，私がオーガナイザーの1人を務めるWMT（Conference on Machine Translation）と呼ばれる機械翻訳の国際会議のコンテストで，中国・テンセントのチームがニューラルネットを用いた自動評価で圧勝する出来事がありました．ところがプロの翻訳者による人手評価では勝てていなかったのです．つまり，彼らは自動採点プログラムに最適化されたシステムを開発してきたということ．

＊5 推論モデル：次の単語を予測するだけでなく，内部で複数の手順を検討し，答えに至るまでの考えを組み立てる（Chain of thought）ように設計されたAI．これにより，それまで苦手とされていた数学の文章題も解けるようになりました．

TencentのHunyuan-MTが自動評価では断トツ1位。でも人手評価は…

English-Japanese									
System Name	LP Supported	Params. (B)	Humeval?	AutoRank ↓	CometKiwi XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET XL ↑
Shy-hunyuan-MT	✓	7	✓	1.0	0.687	82.2	89.6	-5.5	0.592
In2x	?	72	✓	2.3	0.711	78.4	86.3	-5.9	0.575
▲Gemini-2.5-Pro	✓	?	✓	2.4	0.672	83.2	91.2	-5.7	0.55
▲GPT-4.1	✓	?	✓	2.9	0.674	81.8	89.7	-5.9	0.558
Wenyiil	✓	14	✓	2.9	0.682	79.6	88.6	-5.7	0.553
KIKIS	✓	18	✓	3.1	0.678	80.2	85.4	-5.5	0.551
Algharb	✓	14	✓	3.2	0.678	80.8	89.2	-5.8	0.541
CommandA-WMT	✓	111	✓	3.6	0.694	76.5	85.5	-5.8	0.55
▲DeepSeek-V3	?	671	✓	4.6	0.667	80.5	87.7	-6.2	0.531
▲Mistral-Medium	?	?	✓	5.4	0.675	77.6	86.2	-6.4	0.532
GemTrans	✓	27	✓	5.5	0.667	72.3	80.2	-5.5	0.553
▲Claude-4	✓	?	✓	5.7	0.677	78.3	86.3	-6.5	0.516
Yolu	✓	14	✓	5.9	0.697	69.9	77.9	-5.9	0.541
▲ONLINE-B	✓	?	✓	6.1	0.684	72.4	80.0	-6.0	0.527
UvA-MT	✓	12	✓	6.4	0.691	72.5	82.0	-6.3	0.517
bb88	?	?		7.2	0.674	74.6	82.6	-6.5	0.498
▲CommandA	✓	111		7.3	0.674	74.8	82.9	-6.6	0.504
▲Qwen3-235B	✓	235		7.3	0.667	74.3	83.2	-6.4	0.499
Systran	✓	18	✓	7.3	0.703	68.7	77.1	-6.5	0.523
▲Gemma-3-27B	✓	27		7.9	0.666	74.3	82.2	-6.6	0.497
NTTSU	✓	14	✓	8.0	0.676	67.7	74.3	-5.6	0.498
▲TowerPlus-72B[M]	✓	72		8.6	0.671	71.4	80.6	-6.8	0.499
▲Llama-4-Maverick	✓	400		9.1	0.661	71.5	81.1	-6.8	0.487

※AutoRankは，LLMによる評価（GEMBA），正解訳との類似度（XCOMET），原文との類似度（CometKiwi）など複数の異なる自動評価尺度の出力を総合的に加味した順位．1から参加システム数までの値の範囲に正規化されている．テンセントのHunyuan-MTの人手評価は，例えば英中翻訳では2位グループ（2位から4位）だが，中日翻訳は8位グループ（8位から12位）と必ずしも1位ではない．

出典：https://arxiv.org/abs/2508.14909より引用

図3　WMT2025 汎用機械翻訳（GMT）の自動評価

裏を返せば，LLMの自動評価はいまだ人間の能力に達していません．今後の機械翻訳の課題は頑健な自動評価尺度の開発にあるのです（**図3**）．

そこで私たちは，アジア言語の機械翻訳に関するワークショップWAT（Workshop on Asian Translation）における，特許請求項翻訳タスクでの専門家の評価を踏まえて，より精度の高い自動評価プログラムの開発を行っています．人間の評価自体も不安定なので，いかにして安定した評価指標をつくるのかが現在の課題です．

さらに現在，LLMの推論モデルを使った意訳にも取り組んでいます．例えば「井の中の蛙大海を知らず」を英語に訳すのは容易ではありませんが，推論モデルを使って思考過程をLLMにしゃべらせると，意味を汲んだ訳を出せます．これを出発点として，今話題のAIエージェント機能を機械翻訳に取り入れて，専門用語や固有名詞などの正確な翻訳，前後の文脈や文化的背景に即した自然な翻訳ができるよう，自分で判断して外部ツールを呼び出して推論するような機械翻訳エージェントの開発をめざしています．

このように目まぐるしく状況は変わりますが，私自身，諦めずに研究を続けてきたからこそ成果を生み出すことができました．生成AIでかなりのことができるようになってきたものの，特許や法律，医療などの分野では，まだまだ精度向上が求められています．今後はさらに，人間と機械のより良い協調も検討していかなければなりません．いまや最先端の研究にキャッチアップする良いツールもたくさんあるので，それらを使って自己研鑽しながら社会に貢献できる研究をしていってもらいたいと思います．