4 多様な翻訳候補の生成

目的に合う翻訳文を選択できるよう、機械翻訳で多様な翻訳候補を提示

さまざまな機械翻訳エンジンが登場しており、機械翻訳が広く活用されようになった。ただしそれらの機械翻訳を利用すると、翻訳候補文が似通った文章になりやすいという問題がある。NTT コミュニケーション科学基礎研究所（以下、CS 研）ではこの問題の解決策になり得る技術を研究している。

NTT コミュニケーション科学基礎研究所
協創情報研究部
言語知能研究グループ
研究員　**森下 睦**氏

伝えたい内容や文脈により適切な翻訳文は異なる

誰でも無料で利用できる機械翻訳サービスも珍しいものではなくなり、機械翻訳が幅広く使われるようになった。もちろん完璧な翻訳エンジンというものは存在しておらず、いくつもの課題がある。その 1 つが、伝えたい内容や文脈によって適切な翻訳文は異なることへの対応だ。

図 1 に示す“京都は世界的な観光地である”という入力文に対しては、さまざまな翻訳文を考えることができる。どの翻訳文が適切かは目的や状況により異なるため、一概にどれが良いとは言えない。たとえば京都府の広報担当者が一番京都のことを魅力的に伝えられる文を選ぶとしたら、one of the best という表現が使われている最後の文かもしれない。

多様な翻訳文を提示することができれば目的に合う翻訳を人間が選択できるようになり、結果的に翻訳品質を高めることにつながる。

「実現したいのは、パソコンなどで日本語を入力する際にローマ字やかなでの入力に対し複数の日本語の変換候補が出力され、その中から適切なものを選ぶのと同じようなことと言って良いでしょう。」（森下氏）

図 1　伝えたい内容や文脈により適切な翻訳文は異なる

多様な翻訳候補の提示を可能にする上での課題

機械翻訳により複数の翻訳候補を出力すること自体は可能だが、標準的なニューラル機械翻訳モデルは多様な翻訳候補を出力しない。たとえば 5 つの候補を出力させようとしてもほぼ同一の翻訳文が並ぶことになり、図 1 に示すような多様な候補を提示することは難しかった。

「この問題はこれまでにも研究されてきましたが、大きく分けて 2 つの問題がありました。1 つは翻訳モデルが出力する候補が似通ってしまう問題です。あまり似たものばかりでは選びようがありません。もう

1つは多様な候補を出力しようとすると各候補の翻訳精度が低下してしまうという問題です。翻訳精度を保ちつつある程度ばらつきのある多様な翻訳候補を提示することは困難であり、トレードオフの関係にありました。」（森下氏）

翻訳精度と翻訳候補の多様性を両立させる新手法を提案

CS研は奈良先端科学技術大学院大学との共同研究により、翻訳精度を一定程度保持しつつ、多様性の高い翻訳候補を提示する新たな手法を考案した。従来手法と新手法の比較を図2に示す。多様性を示すDPという指標と、翻訳精度を示すBLEUという指標は、どちらも大きい数字ほど良い結果であることを意味する。

多様性が51.3から67.7と明らかに向上しつつ、翻訳候補の精度もほぼ同じ、むしろごくわずかながら良いということがわかる。

2023年6月1～2日に開催されたCS研の「オープンハウス2023」では、この新手法による翻訳候補を体験できるデモ展示が行われた。ブラウザ上で日本語を入力すると多様な翻訳候補が表示されるというものであった。図1に示した翻訳候補の例もこのデモと同じ仕組みで出力されている。

	多様性（DP）↑	翻訳候補の精度（BLEU）↑
従来の多様性を考慮する手法	51.3	37.9
提案法	67.7	38.4

（DP: 出力された候補文にどれくらい異なった単語/フレーズが含まれているかを示す指標）

図2　提案手法の効果

以下、この新手法の特徴について説明していく。

従来の「検索に基づく翻訳手法」を利用

ベースとなっているのがkNN（k-nearest neighbor）機械翻訳と呼ばれる検索に基づく翻訳手法だ。一般的な機械翻訳はニューラルネットワークによる翻訳モデルを利用するのだが、kNNではそれに加えて膨大なデータベースを利用する点が大きな特徴となっている。

このデータベースは日英翻訳向けであれば日本語と英語の対訳がまとめられた、いわゆる対訳コーパスを基に構築される。入力文と類似する事例をこのデータベースから検索し機械翻訳に活用することで、翻訳精度が向上することが知られている。

「このデータベースにはニューラルネットワークの内部状態とその次に出力すべき単語のペアが大量に蓄積されています。図3の例のような“このヌマガメはおよそ70歳だ”という入力に対しThisと出力した次にはterrapinが来ることを示す情報です。この情報が示すベクトル空間には単語の数だけ点があります。何億のような膨大な数です。」（森下氏）

図3右に示すように“このカメは主に何を食べますか？”という入力に対して検索を行うとWhat does thisの後にカメに相当する単語の候補としてturtleやturtles、slider（アカミミガメ）、tortoise（リクガメ）、terrapin（ヌマガメ）などが該当する。検索の範囲を絞り込むことにより日本語のカメそのものに最も近いturtleとturtlesを出力するという仕組みだ。kNNという名前は「k

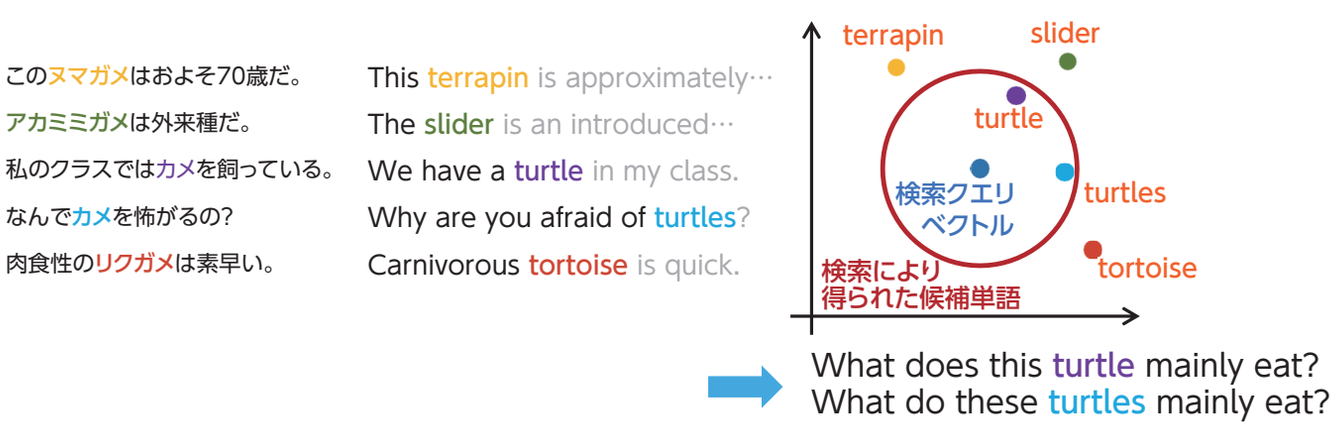

図3　従来の検索に基づく翻訳手法（kNN機械翻訳）

個の近傍」を取得することに由来している。図３右の例では k が２となっている。

摂動を加えた kNN 機械翻訳で多様な候補の提示を可能に

森下氏が取り組んだのは、翻訳精度の向上に役立つ kNN 機械翻訳を活用しつつ、多様な翻訳候補を得られるようにすることであった。そのため図４に示すように検索の範囲を広げ、さらに翻訳候補として活用するデータをランダムに選択するようにした。

さらに従来手法で広く使われている Beam Search という生成手法の替わりに、Divers Beam Search（以下、DBS）という生成手法を活用している。翻訳候補の多様性が低下する原因の一つが生成のしかたにあったためだ。DBS は多様化に役立つものの、kNN 機械翻訳では活用されていなかった。また少なくとも実用レベルの機械翻訳ソフトウェアやサービスでの利用例はないと考えられる。

DBS を活用することに加え、単にばらつきのある翻訳候補を得るのではなくより適切な単語群の中から多様な翻訳候補を得ることができるよう工夫している。

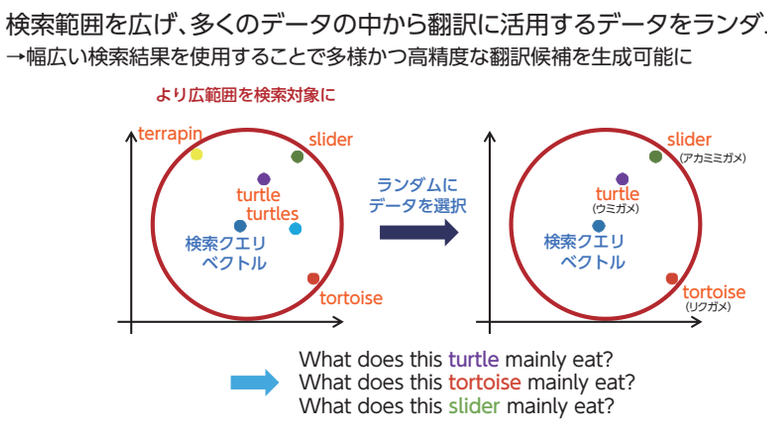

図４　検索にランダム性を加えた翻訳手法

「従来の検索に基づく翻訳手法ではカメに対して turtle と turtles のように似通った翻訳候補ばかり選択されやすかったのに対し、検索にランダム性を加えることで図４の例のように slider や tortoise といったカメを意味する他の単語も候補にできるようになりました。」（森下氏）

従来手法と新手法を多様性と翻訳精度による評価で比較

新手法の有効性を評価した結果を図５に示す。点線で区切られた上半分はニューラル機械翻訳モデルによる従来の一般的な機械翻訳、翻訳候補の検索に DBS を用いる機械翻訳、kNN 機械翻訳という３つの手法、下半分は CS 研が提案した新たな手法であり、DBS と kNN を組み合わせた機械翻訳、さらに検索範囲を広げてランダムに k 個の単語を選択する機械翻訳の２つとなっている。

Method	DP (%)	BLEU@20
一般的な機械翻訳	32.1	51.4
DBS	41.3	47.6
kNN機械翻訳	31.7	52.2
DBS ＋ kNN機械翻訳	41.7	48.7
＋ ランダムにk個選択	42.2	48.6

図５　従来手法と新手法の比較

DP は多様性、BLEU@20 は翻訳候補を 20 個提示しその中から最適な翻訳結果を選択した場合の翻訳精度を示している。

DBS と kNN 機械翻訳を組み合わせた手法は多様性と翻訳精度を一定程度両立できており、ランダムに k 個を選択することでより多様性を高める効果もあることがわかる。

大規模日英対訳コーパス「JParaCrawl」を活用

本稿で紹介したのは翻訳モデルに関する研究成果だが、機械翻訳では対訳データも重要だ。kNN 機械翻訳の翻訳精度もデータベースに大きく左右される。CS 研ではこのデータベースの構築に、独自に作成した最大規模の日英対訳コーパスを使用している。森下氏自身が、質の良い対訳コーパスを得るためいかに対訳

データを収集するかという研究テーマにも取り組んでいる。

「対訳データの収集自体は以前から取り組んでいる研究テーマです。インターネットを通じて対訳データを集めるため、どの Web サイトに対訳データがあるのか、日本語と英語の対がどのページにあるのか、について知ることから取り組んでいます。2020 年からは研究成果を大規模日英対訳コーパス JParaCrawl として公開しています。これまでに数回アップデートを行っており、現在の規模は 2200 万文対を越えています。多様な翻訳候補を提示するための研究においても JParaCrawl と同等のデータを使用し、kNN 機械翻訳向けのデータベースを構築しました。」（森下氏）

より翻訳精度と多様性を高めるための研究を続ける

引き続き研究を進めたい技術的な課題の 1 つとして、森下氏は「検索で見つけ出した多数の単語をいかに選択するか」を挙げている。現在は広めの検索範囲から見つかる多数の単語のうち、いくつかの単語をランダムに選択している。

「ランダムに選択する現在の方式では、検索範囲を広げた際にノイズのような翻訳精度の低下につながる単語も含まれてしまう可能性があります。そのような単語を選ばないようにするにはどうすれば良いか、もしくはより多様性に寄与するよう選ぶにはどうすれば良いか、といった技術的な課題に取り組みたいと考えています。」（森下氏）

ニーズに合わせた翻訳候補の出力や並べ方も課題の 1 つ

より使いやすいよう翻訳候補を提示することも課題の 1 つだ。どの翻訳候補がより適しているかをすぐに判断できるのは、優れた語学力のある人に限られる。そこで最も適切な翻訳から順に翻訳候補を並べて提示することにも取り組む考えであるという。実現すれば高い語学力を持たない人でもより良い翻訳文を選択しやすくなる。

この優先順位付けは、メールを返信するための文章なのか、その相手は誰なのか、また特許情報のような特殊な翻訳をしているのか、といったように目的によって適切な解が異なる。

「より魅力的な文章に見えるようにしたい、というニーズもあるかもしれません。そのようなユーザーのニーズに応じた翻訳候補の出力、また並べ方、見せ方といった部分にまだ課題があると思っています。」（森下氏）

この研究課題は目的に合わせて翻訳する、すなわちドメイン適応を行うということと関係が深い。データベースを活用する kNN 機械翻訳はもともとドメイン適応に有利な翻訳手法と考えられており、研究の進展を期待できる。

翻訳以外への応用の可能性も

本稿で紹介した新手法はあくまでも翻訳を目的としたものだが、森下氏はそれ以外の分野への応用も可能ではないかと考えている。

「対話の中で使われる文は多様であり、翻訳よりもさらに多くの選択肢があり得ます。たとえばチャットボットのような自動応答の出力に本技術を応用できるかもしれません。」（森下氏）

CS 研の強みを活かし将来を見据えた研究を続ける

オープンハウス 2023 におけるデモ展示では、さまざまな来場者から「これは使える」といった好反応を得られた。一見するとこのままでも使い道があるような仕上がりでもあるが、森下氏によれば今すぐ実用化するのは難しいという。ただしそう遠くない将来の実用化が可能と考えているとして、同氏は次のように述べている。

「検索に基づく翻訳手法は単語毎に検索を行うため、検索を行わない一般的な機械翻訳と比較すると原理的に処理に時間がかかります。ですが検索に基づく翻訳手法の処理速度を向上させる研究も行われており、遠くない将来に充分な速度で処理できるようになると考えています。ほかにもいくつか障壁はあるのですが、遠い将来ではなく何年後のような期間で実用化が可能になることを想定し、研究に取り組んでいます。

CS 研には自然言語処理を研究テーマとする研究者が多数揃っており、国内では有数の研究体制と言えます。また、より良い機械翻訳を実現するためには翻訳モデルと対訳データのどちらもが重要です。その両方の研究に力を入れていることも、我々の強みの 1 つです。今後もこうした強みを活かしながら研究に取り組む考えです。」