金融×IT 対談

サウラブ・バジ氏

Cohere

CTO（最高技術責任者）

Profile/Saurabh Baji

Cohereにて、機械学習、エンジニアリング、プロダクトマネジメントを統括。これ以前は、AWS、Unity、Quantcastにて20年以上にわたりリーダー職を歴任し、ビッグデータ、ロボティクス、科学計算、物流、サプライチェーン最適化など、さまざまな分野でAIを活用した数十億ドル規模の製品の開発と運用をグローバルに推進。

向井 佳子

株式会社野村総合研究所

金融AIプラットフォーム推進部長

Profile/Keiko Mukai

1997年 大手IT企業入社。99年 野村総合研究所入社。大手証券会社向けの営業店フロントシステム企画・開発や、オンライントレードシステムの事業を担当。2016年 リテールプロジェクト推進室長を経て、2025年 現職に就任。金融機関向けのAI活用プラットフォームの企画に取組中。

金融業界では大規模言語モデル（LLM）等のAI技術を業務に活用する機運が高まっているが、安全性の確保など課題も多い。企業の高度な要求に応えるLLMのソリューションとはどのようなものか。エンタープライズ向けに最先端のLLMを提供するAI企業、Cohere社（米・カナダ）で技術開発をリードするバジ・サウラブ氏に語っていただいた。

Cohereの生成AIの強み

向井　御社は、企業向けに大規模言語モデル（LLM）を提供していらっしゃいます。まず事業内容や主力製品について概要を教えていただけますか。

バジ　Cohereは、AIモデルと包括的なソリューションの両方をエンタープライズ向けに提供している会社です。背景には「生成AIは単なる便利ツールではなく、企業の抱える課題を解決するものだ」というわれわれの哲学があります。

近年、生成AIの分野では、魔法のような技術革新があり、それまでのAIが上手くできなかったタスクを容易に実現できるようになりました。しかし、その後もモデルは進化し続けているにも関わらず、多くの企業では実際のビジネスに上手く活用できていません。そうした状況を目の当たりにして、われわれはモデル開発に加えてAIを実ビジネスに実装するためのソリューションにも注力するようになりました。

弊社では、顧客ニーズに合わせて複数の異なる生成AIモデル（Command A、R、R+、R7B）

を提供しています。これらのモデルは多言語対応で、テキスト生成・要約、企業内情報を活用したFAQの実装に使われるRAG（検索拡張生成）機能等を備えています。

RAGには、その性能を決定づける3つの重要な要素があります。

1つ目は検索・抽出の精度です。ユーザーの質問に対し、関連性の高い情報をデータソースからどれだけ正確に見つけ出せるかです。

2つ目は取得した情報の関連性評価です。検索で抽出された複数の情報の中から、ユーザーの質問の意図に合致する、より重要度の高い情報を選び抜く能力が重要です。

3つ目はLLMのRAGプロセス実行能力です。生成AIのモデルが、より的確な情報を得るために検索クエリを分解・再生成（サブクエリ生成）したり、複数の異なるデータソースからの情報を統合して扱ったりできることが重要です。

弊社では、情報検索と抽出のためのモデルとして、検索精度を高めるための技術（エンベディング）と検索結果を関連性の高い順に並べ替える技術（リランキング）を提供しています。これらは、企業が保有する様々なデータから正確かつ安全に関連情報を抽出するのを助けるものです。企業は膨大なデータを保有しているため、これらのモデルの品質はRAGの性能に大きく作用します。

向井　今おっしゃったリランキングモデルは御社の強みだと理解しています。

バジ　弊社のリランキングモデルは世界最高レベルにあります。リランキングは他社に先駆けて弊社が最初に提供を開始しました。多くの競合他社はこのリランキング機能を提供していません。

リランキングを使うことで企業はAIの利用コストを節約することができます。企業は、大量のドキュメントを保有しており、それらのすべてを入力データとして生成モデルで処理しようとすると非常に高いコストがかかります。しかし、エンベディングでセマンティック検索（質問の意味や文脈に基づく検索）を行い、さらにその結果に対してリランキングを実行すれば、最も関連性の高いデータのみを渡すことができます。

向井　生成モデルに大量のデータを処理させるとコストが高くなるので、それを減らすことができるわけですね。

バジ　そうです。例えるなら、探している答えが本のどこか1ページにあると分かっているのに、本全体を探す必要はないということです。エンベディングを使うと関連する情報がいくつかに絞られ、さらにリランキングを行うと、最も関連するページがわかり、そのページだけを生成モデルに送ることができるわけです。

それからもう1つ、弊社の生成モデルが得意としているのが、AIの回答にサイテーション（引用情報）を付けて出力することです。サイテーションをつけることで、AIがどの情報に基づいて回答を生成したのかを後から確認することができます。Cohereでは1年半ほど前からサイテーション機能を提供しています。

向井　生成された回答のそれぞれの文が、どのドキュメントのどのページに基づいたものかわかるということでしょうか。

バジ　その通りです。生成AIには、事実に基づかない情報を生成する、という課題があります。ハルシネーションと言われているもので、企業にとっては大きな懸念事項の一つです。サイテーションがあれば、情報が存在すること、さらにはどこに情報があるかが正確にわかり、ハルシネーションを抑制できます。情報を見つけられなければ、モデルは「答えはわかりません」と応答します。誤った答えを回答するよりずっとよいでしょう。

企業の高度なニーズにどう応えるか

向井　日本の金融機関は昔からのドキュメントを大量に保有しており、中には制度変更などによって古くなってしまった情報も含まれています。人間であればどれが古い情報かすぐに判断できますが、AIは古い情報でも新しい情報と同等に扱って高くランク付けしてしまう可能性がありそうです。御社のリランキングモデルでは、どのような基準でランク付けが行われているのでしょうか。

バジ　リランキングでは、どの情報が最も関連性が高いかを判断するため、様々な「関連性要素」を考慮してランク付けを行っています。中でも、情報の新しさを示す時間的要素は重要な要素の一つです。ご指摘のように、古いドキュメントは、たとえ検索で完全一致したとしても、現在の状況と関連しない可能性が高いためです。

なお、古い情報を排除するために、

クエリ自体に時間的制約を含めることも可能です。昨年12月にリリースした最新のリランキングモデルは、クエリに基づいて推論する能力が向上しており、例えば「1か月以内の結果のみ」といった制約を理解することができます。

向井　リランキングを実行する際、どの要素を重視して情報の優先度を並び替えるべきかといった基準をユーザーが意図的に指定することはできるのでしょうか、それともユーザーのクエリに基づいて、AIが自動的に要素を決定するのでしょうか。

バジ　われわれはリランキングモデルが多くのユースケースに対応できるように、さまざまな要素を学習させています。情報の新しさ、内容の関連性などの要素を総合的に判断して、検索結果をランク付けします。基本的にはAIが自動で最適な判断を行いますが、お客様のニーズに応じて、特定の要素を重視するように調整することも可能です。

向井　企業向けのLLMにおいては、セキュリティやプライバシーの問題も非常に重要です。御社では何か特別な対応をされていますか。

バジ　弊社では、お客様のデータがどこに置かれていても、われわれのモデルやソリューションを展開できるようにしています。競合他社と同じようにSaaSソリューションを提供していますが、OCI、AWS、Azureなどの主要なクラウドプロバイダーのマネージドサービス上でも利用できます。さらに全く同じ技術を仮想プライベートクラウドやオンプレミス環境、エアギャップ環境でも展開可能です。

ですから、高い機密性を要するユースケースであっても、お客様はデータ漏洩やセキュリティを懸念することなく、弊社のモデルを利用することができます。

生成AIの様々なリスクへの対応

向井　欧州と米国では、AI規制に対するアプローチがかなり異なると理解しています。欧州はAIのリスクを懸念し規制を強化する傾向にあります。一方、米国は技術革新を重視し自由な開発を促進する傾向にあります。御社のAIのリスクに対する見解は、どちらの立場に近いと思われますか。

バジ　われわれとしては、セキュリティ、プライバシーといった目の前にある現実的なAIのリスクについてはお客様が懸念される問題だと考えています。ですから、お客様がそうしたリスクを抑制するための支援は惜しみません。これに対して、AIが人類の存在を脅かすといった長期的な脅威については、実際にはなかなか発生しないのではないかと思っています。各業界でも政府機関でもそうしたリスクを注視しており、現実に懸念すべき問題が生じたら対処していくことができると信じています。

弊社は、技術を過度に制約するとその恩恵を十分に受けることができなくなると考えているため、米国のアプローチに近いと思います。とはいえ、地域によってリスクに対する認識が異なるのも認識していますので、お客様と協力して各国・地域の規制に対して適切な対応を行っています。

向井　日本の金融機関では、AIによる人権侵害のリスクを非常に気にしています。こうした問題についてはどのように取り組んでいらっしゃいますか。

バジ　人権に関する質問など物議を醸す可能性のある質問に対してAIがどう回答すべきか、といった問題は非常に大きな問題です。ほとんどの場合、企業は可能な限り安全な回答を求めているため、当社のモデルは、初期設定では非常に保守的な回答をするように設計されています。しかし、レッドチーミング（脆弱性診断）のような、あえてリスクのある回答を求めるケースでは、安全モードを解除することで、柔軟に対応できます。

向井　日本では2022年から経済安全保障推進法が施行され、金融機関の間では経済安全保障面の各種対応の重要性が高まっています。御社では経済安全保障関連の法規制に絡むコンプライアンスについても意識されていらっしゃいますか。

バジ　弊社は、そうした問題に対して適切に予防措置を講じております。

オープンソースのソフトウェアは利用していますが、弊社のモデルはすべて自社で開発したもので、モデ

ルの学習でも細心の注意を払っています。

金融機関は生成AIに何を求めているか

向井　御社は多くの金融機関とLLMを導入するために協業されています。そうした経験から、どのようなところが金融業界特有のニーズと思われますか。

バジ　大きく2つあると考えています。

1つ目は、どの金融機関にとっても、情報セキュリティと顧客情報の保護が最優先事項であるということです。お客様は生成AIを活用して生産性や顧客サービスを向上したいと考えてはいますが、それは情報セキュリティとプライバシー保護が完全に確保された環境で実現されなければなりません。

2つ目は、汎用的なAIモデルだけでは、金融業界特有の複雑な課題を解決することは難しく、業界や個々の金融機関のニーズに合わせたソリューションが必要だということです。中でも、あるカナダの有力銀行のお客様との協業では、金融業界に対応したエンドツーエンドのソリューションに対して大きなニーズがあることを学びました。

向井　「エンドツーエンドのソリューション」とは、具体的にどのようなものなのでしょうか。

バジ　エンドユーザーが技術的な手間をかけることなく、すぐに使い始められるソリューション、ということです。エンドユーザーがデータソースの接続やドキュメントのアップロードを気にしたり、データやプライバシーの問題を心配したりする必要はない、ということです。

こうしたニーズに応えるため、弊社では、業種ごとにカスタマイズできる、企業向けのセキュアなAIプラットフォーム「North」を開発しました。このカナダの銀行とは、金融業界に特化したソリューションとして「North for Banking」を開発しているところです。

向井　Northでは、どのようなことが実現できるのですか。

バジ　Northは、エージェント的なワークフローを備えており、企業の様々なデータソースに接続することで、社内の法務、財務、人事に関する質問にも、顧客向けの製品の質問にも回答することができます。Northを利用するのに機械学習のエンジニアは必ずしも必要ありません。AI人材が不足している企業にとってはこうしたエンドツーエンドソリューションは大きな意味を持つでしょう。

向井　AI人材の不足をどう解決するかは企業が直面している課題の一つだと思いますので、AIエンジニアでなくともAIの実装ができるようになることは頼もしいですね。

バジ　生成AIは金融業界においてより重要な役割を果たすようになると確信しています。多くのユースケースで高いROI（投資対効果）を生み出せる可能性があると考えています。生成AIによって、より正確、迅速、包括的に回答を得られるようになり、金融機関もそのお客様もどちらも恩恵を得ることができるでしょう。

金融機関のユースケースは幅広く、財務諸表の作成や分析、ポートフォリオ管理、不正検知、リスク評価と管理、顧客サポートなど様々です。早くからLLMに投資してきたお客様では、すでに成果が出始めています。2023年にはいくつかの金融機関がPoC（概念実証）を開始し、2024年には、その多くが本番稼働に移行しています。

向井　早く始めるほど大きな成果が見込めそうですね。

バジ　早く始めればそれだけ早くより多くのノウハウを蓄積し、競争優位性を確立できるのではないでしょうか。

私はお客様に「Think big, but start small（大きく考え、小さく始める）」と助言しています。適切な規模のユースケースで素早く課題を解決できれば、そこで自信を深め、より複雑なユースケースに取り組むことができるでしょう。

向井　かつては大企業がスタートアップを選ぶ時代でしたが、今は御社のようなスタートアップがパートナーを選ぶ時代です。NRIも努力しますので、ぜひ一緒に協力して取り組んでいければ幸いです。

本日は貴重なお話をありがとうございました。

（文中敬称略）