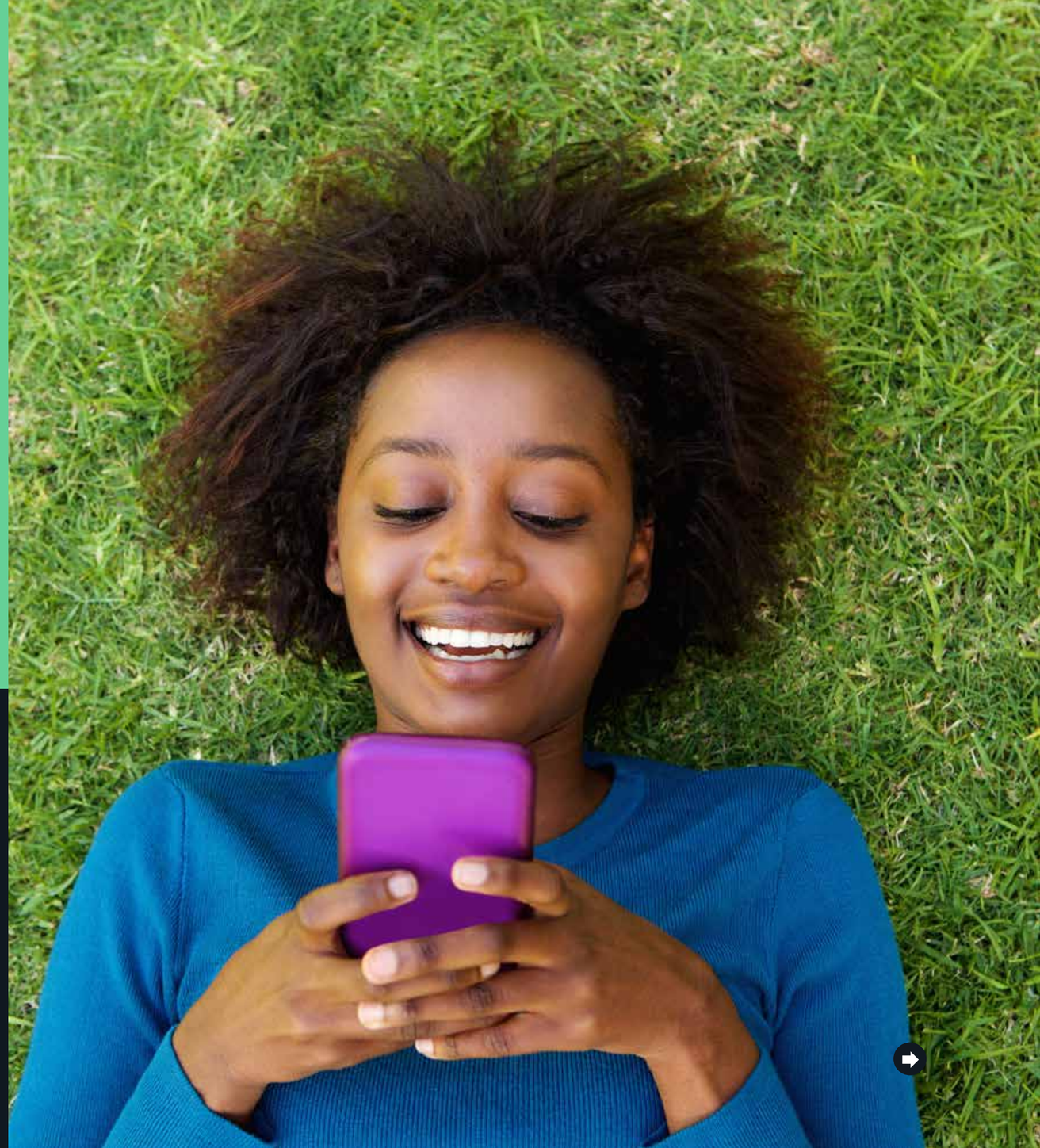
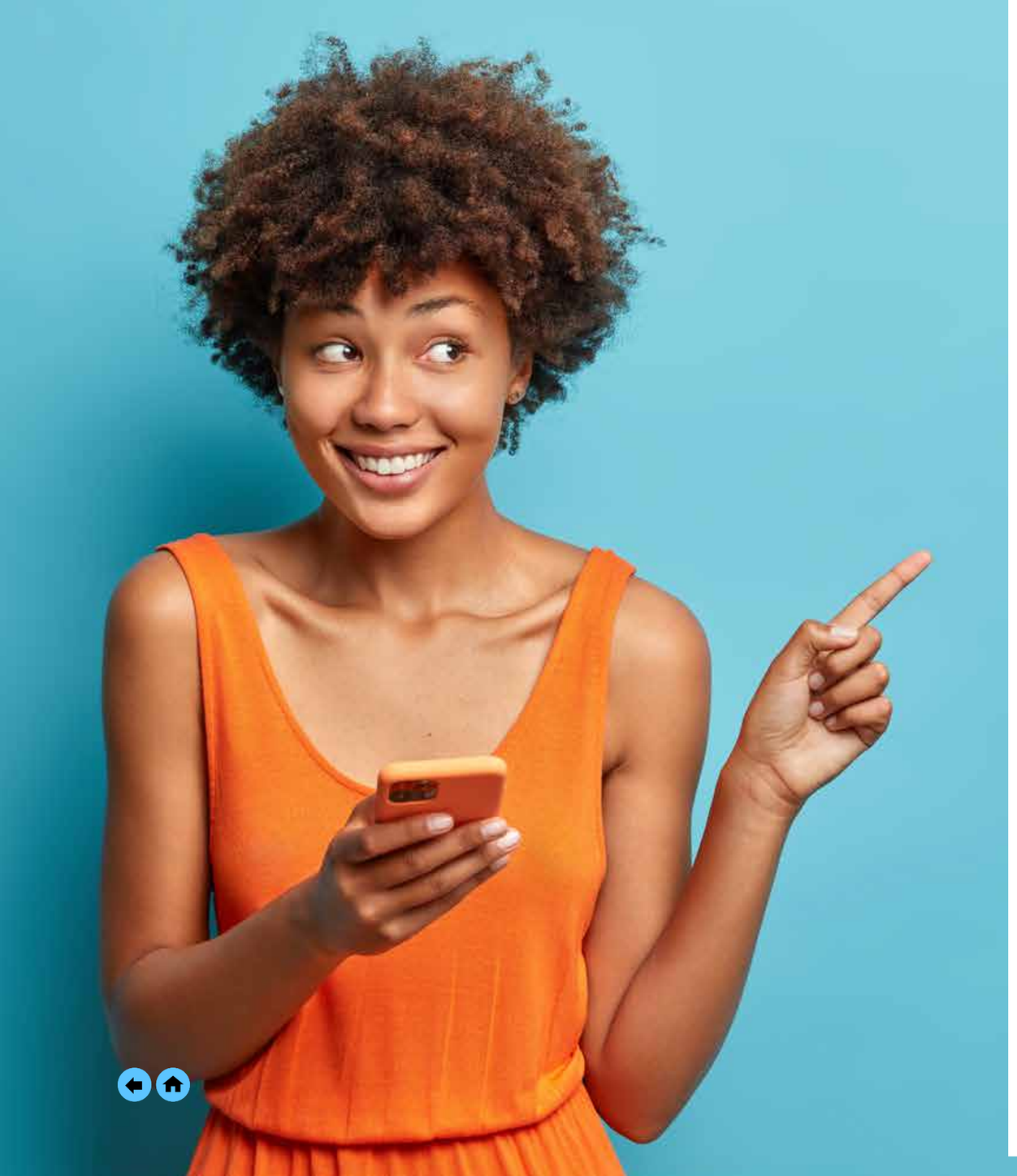
Eブック

より自然な言語処理が応用分野を広げる

会話型AIのためのデータ インフラの構築

NetApp





目次

限られた分野での応用から
より広範囲な活用へ

自然言語処理（NLP）は会話型AIとして知られ、
「トーキング ロボット」とも呼ばれています。

呼び名はどうであれ、会話型AIシステムは人間のように話し、文脈を理解し、インテリジェントな応答を返します。これはすべて、ディープ ラーニングの飛躍的進歩により、AIシステムが機械的ではなく、より自然な反応をするようになったおかげです。

ディープ ラーニングによって、AIが親しみやすくなるだけでなく、バックエンドにある言語学とルールベースの手法に関して人間が深い知識を持つ必要がなくなります。そのため、込み入った専門用語を使う業界（金融サービス、ヘルスケアや生命科学、政府機関、自動車、製造業、小売業など）では、自然言語処理ソリューションをスムーズに導入することができます。

会話の質を高める鍵はデータ

このようなAIモデルは処理される情報が膨大で複雑さを極め、思考と同じスピードで大量のデータを移動する必要があります。これを成功させるためには、自然言語処理のためのインフラで次のことが可能でなければなりません。

1. ボリュームを増やす
2. パイプの詰まりをなくす
3. 瞬時に応答する

自然言語処理：チャットボットに留まらない活用

スマート アシスタント、検索エンジン、入力予測などに応用される自然言語処理は、新たなグローバル言語です。あらゆるところで活用され、ときには、思いもよらないところで活用されていることもあります。

信用力の評価

自然言語処理は、位置情報、ソーシャル メディア アクティビティ、閲覧行動、ピア ネットワークなどのデータに基づいて信用度スコアを生成するために使用できます。

治験のマッチング

患者に治験への参加を促すのは難しいことですが、その理由のほとんどは、治験があることが知られていないからです。自然言語処理を応用すれば、研究者や製薬会社は患者と治験を自動的にマッチングさせることができます。

法律の執行

警察は、自然言語処理を使って犯罪の動機を特定することで、治安を維持し暴力を抑制すると同時に、取り締まりを明確化し、即応性を高めています。

車両整備

自然言語処理を活用すれば、ドライバーが車を最高の状態に保つのは簡単です。分厚い取扱説明書を読む必要はなく、オーナーは車に「どの警告灯が点灯している？」「ヒューズの交換方法を教えて」と問いかけるだけです。

航空機の修理

自然言語処理は、整備士が膨大なサービス マニュアルの情報を全体的に把握し、パイロットから報告された問題を深く理解するのに役立っています。

1. ボリュームを増やす

自然言語処理を適切に活用するためには、信じられないほど大量のデータが必要です。これまでに使われたすべての言葉に相当する量だと思えば、それがほぼ正解でしょう。

自然言語処理は、音声入力を処理して理解し、膨大なデータのライブラリを参照し、人間が理解できる応答を数ミリ秒以内に生み出す必要があります。

多くの規則と例外がある人間の言語の複雑さを考えると、この要件は非常に困難ですが、慣用句や皮肉、ユーモアが持つニュアンスを考慮すると、さらに難しくなります。また、業界特有のモデルにより、特定の分野、企業、製品に関する特定の情報が必要になることもあります。

このような理由により、会話型AIモデルの規模は、パラメータ数が数百万個から数十億個にもなります。一般に、データが多いほどモデルの精度は向上します。この規模のモデルのトレーニングには、数週間ものコンピューティング時間がかかる場合があり、機械学習とディープ ラーニングにおけるトップクラスのフレームワークが必要です。

Google翻訳

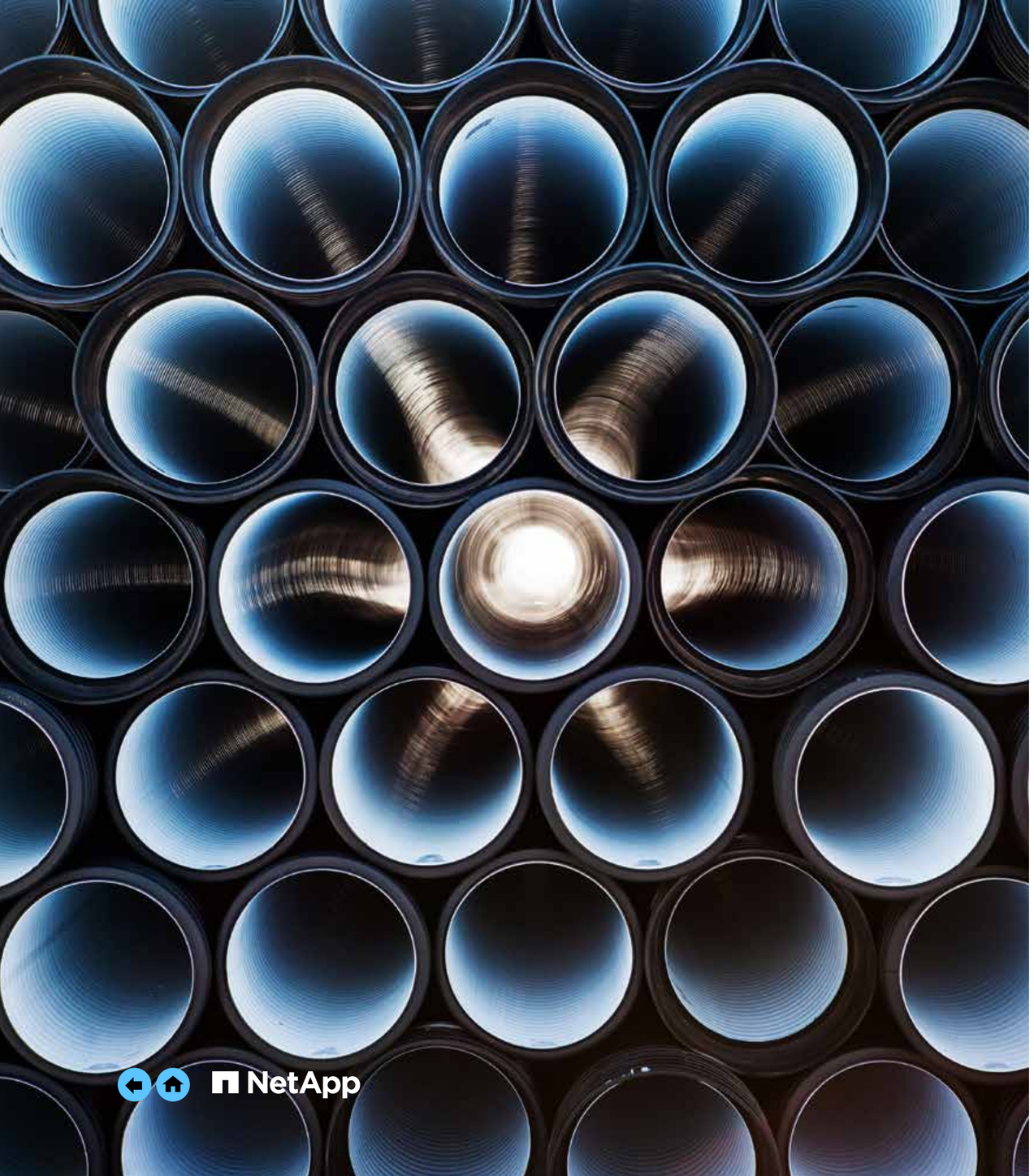
Google翻訳は100を超える言語に対応していますが、トレーニング用のコーパス（ソース データセット）が限定的である言語については、翻訳およびモデルのトレーニングのテストと改善にクラウドソーシングを活用しています。Google翻訳は毎日1,400億語を処理しています。これは人間の翻訳者7,000万人に相当する仕事量で、それを毎日処理しているのです。

Google BERT

Google BERTは著名な自然言語処理モデルであり、3億4,000万個ものパラメータがあります。BERTは、電話自動応答アルゴリズムなどのトランザクション型音声インターフェイスを超えて真の会話型を実現した点で、画期的な自然言語処理と言えます。テキストの読み上げや質問への回答を、きわめて高い精度で実行できます。

BioMegatron

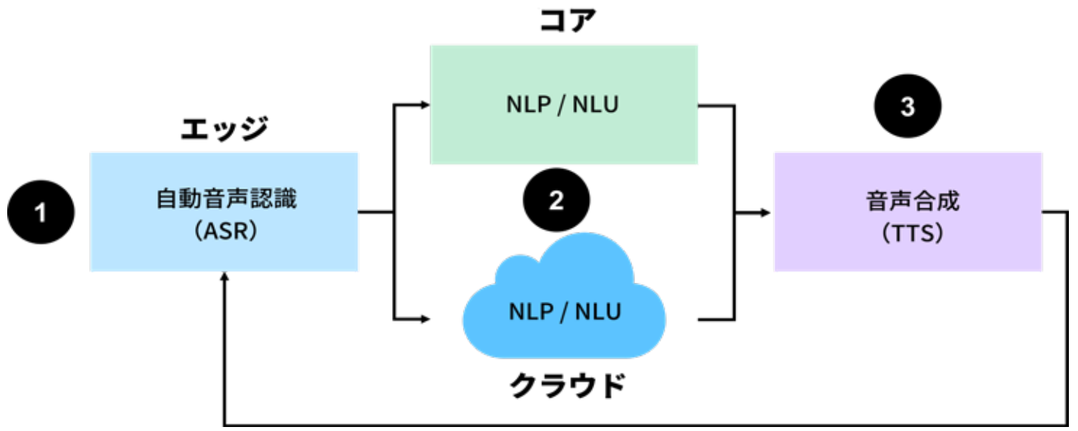
BioMegatronは、これまでにトレーニングされた中でも最大規模の、生物医学分野におけるTransformerベース言語モデルです。パラメータのバリアントは最大で12億個に上ります。生物医学に関する論文の要約と全文のリポジトリであるPubMedから取得された61億語を用いてトレーニングされました。



2. パイプの詰まりをなくす

高速で効果的な自然言語処理には、データの取り込みと認識から音声合成に至るまで、エコシステム全体にわたるデータ パイプラインが必要です。リアルタイムの言語処理を行うためには、パイプラインの各ステップを迅速かつスムーズにデータが通過できなければなりません。

一般的な自然言語処理パイプラインは3つの段階で構成されます。



最先端の自然言語処理インフラでは、毎日、数千カ所のエッジ ロケーションでテラバイト単位のデータが収集されています。このデータへのアクセスが、サイロ化されたインフラによって制限されていると、ディープ ラーニングでデータを深く掘り下げることができません。

3. 瞬時に応答する

AIが人間の発話を真似るためには、人間の脳と同じかそれ以上のスピードで動作しなければなりません。モデルの規模が大きいほど、ユーザが質問してからAIが応答するまでのタイム ラグが長くなります。自然な会話に聞こえるためには、すべての演算が300ミリ秒の間に実行される必要があります。

このプロセスには複数のステップがあります。

1. ユーザの言葉をテキストに変換する
2. テキストの意味を理解する
3. 文脈に応じた最適な応答を検索する
4. 応答を音声で返す

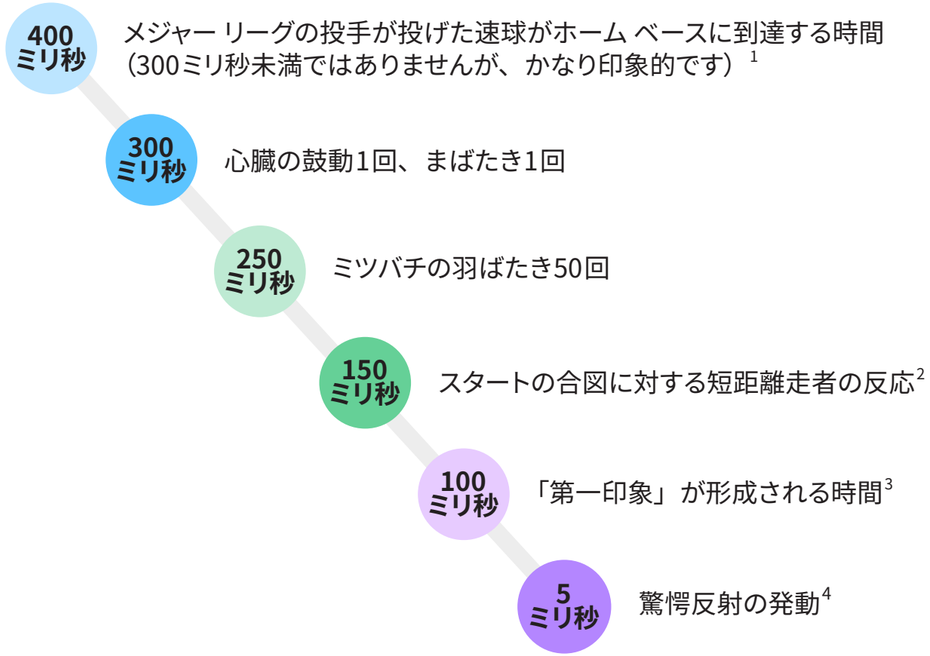
これほど厳しいレイテンシ要件がある状況では、多くの場合、会話型AIの開発者は妥協せざるを得ません。質の高い複雑なモデルでは、迅速に結果を出すものの微妙なニュアンスに欠ける比較的小規模な言語処理モデルよりも、処理に時間がかかる場合があります。

緊張している求職者のように、音声アシスタントが会話中に「私がお調べしましょう」などと発言して時間稼ぎをしたり、気まずい沈黙をごまかすために意味のない電子音を発したりするかもしれません。しかし、自然言語処理の目標である理想的な会話型AIとは、相手の質問を正確に理解できる程度に洗練され、シームレスな自然言語ですばやく応答できる高速さを備えたものです。

人が話すスピード

自然言語処理がリアルタイムに応答を返すためには、一般に、レイテンシを300ミリ秒（0.3秒）未満に抑える必要があります。これはきわめて高速です。

300ミリ秒以内に完了することを挙げてみましょう。



お客様のニーズに応えるネットアップ

NVIDIA DGXシステムとネットアップのクラウド対応オールフラッシュストレージ システムを搭載したNetApp® ONTAP® AIでは、膨大な量の最先端の言語モデルに学習を施し、最適化することで、推論を迅速に実行できます。ネットアップを基盤とするデータ ファブリックは、エッジからコア、クラウドまでのAIデータ パイプライン全体でシンプルなデータ管理を実現します。

- **ネットアップのAIソリューション**は、ボトルネックを解消することで、より効率的なデータ収集、AIワークロードの高速化、スムーズなクラウド統合を実現します。
- **ネットアップの統合データ管理ソリューション**は、ハイブリッド マルチクラウド環境全体にわたる、対費用効果の高いシームレスなデータ移動に対応しています。
- ネットアップは、**ワールドクラスのパートナー エコシステム**を通じて、AIのリーダー企業、パートナー、システム インテグレータ、ソフトウェアとハードウェアのプロバイダ、クラウド パートナー各社と完全な技術統合も果たしており、お客様のビジネス目標の達成を支援するスマートでパワフルな信頼性の高いAIソリューションを実現します。
- **ネットアップ プロフェッショナル サービス**では、複雑さを軽減し、AIのビジネス チャンスと成功を拡大するために必要な専門知識を提供します。

そして何よりも、ネットアップは、IDC MarketScapeにより全世界のスケールアウト ファイルベース ストレージにおけるリーダーに位置付けられています[5]。コンピュータ ビジョンのワークロードは、お察しのとおりスケールアウトかつファイルベースであるため、これは重要なポイントです。

データ サイエンティストを納得させる

5倍

AIパイプラインを通じて
5倍のデータを処理

60秒未満

数時間から
数日かかっていた
データセットのコピーを
数秒で実行

約20分

Ansibleの統合により、
AIインフラを
約20分で構成

NetApp Retail Assistant：成功の青写真

ネットアップとNVIDIAは、会話型AIサービスを構築するためのエンドツーエンドのフレームワークであるNVIDIA Jarvisを使用して、バーチャルの小売業アシスタントを構築しました。このアシスタントは、音声またはテキストによる入力を受け付け、weatherstack API、Yelp Fusion API、eBay Python SDKに接続することによって、天気、特定の地点、在庫品の価格設定に関する質問に答えます。詳細はこちら

NetApp Retail Assistant（NARA）は以下で構築されています。

- **NVIDIA Jarvis：**Jarvisは、レイテンシを低く抑えるように最適化されたエンドツーエンドのディープ ラーニング パイプラインを使用して、GPUによって高速化されたサービスを会話型AI向けに提供します。
- **NetApp ONTAP AI：**この実績のあるアーキテクチャは、NVIDIA DGXシステムとネットアップのオールフラッシュ ストレージを組み合わせています。ONTAP AIは、データ フローを確実に合理化し、レイテンシ要件を超えることなく、複雑な会話モデルをトレーニングして実行します。
- **NVIDIA NeMo：**NeMoは、GPUによって高速化された会話型AIモデルの構築、トレーニング、微調整を行うためのPythonツールキットであり、リアルタイム自動音声認識（ASR）、自然言語処理（NLP）、テキスト音声合成（TTS）といったアプリケーションを含め、使いやすいAPIを備えたモデルを構築できます。

自然言語処理を推進するには

ネットアップにお任せください。まず何から始めましょう。森の生きものとの会話？リスに言葉を教えることはできませんが、自然言語処理に適したAIインフラを構築する方法はお伝えできます。

ネットアップのAIソリューションについて、詳しくはこちらをご参照ください。

- ネットアップのAI
- ONTAP AI
- NLP向けネットアップ ソリューション

ご不明な点がある場合は、ネットアップのAIソリューションのスペシャリストがご案内いたします。

1. O’Neill, Shane、『Real-time bidding: What happens in 200 milliseconds?』（Nanigans）
2. Welsh, Tim、「Exactly how long does it take to think a thought?」（The Christian Science Monitor）、2015年7月1日
3. Wargo, Eric、「How Many Seconds to a First Impression?」（Association for Psychological Science）、2006年7月1日
4. Wise, Jeff、「What Is the Speed of Thought?」（New York Magazine）、2016年12月19日
5. Potnis, Amita、『IDC MarketScape: Worldwide Scale-Out File-Based Storage 2019 Vendor Assessment』（IDC）、2019年12月

ネットアップについて
ジェネラリストが多い世界で、ネットアップはスペシャリストとしての存在感を示しています。お客様がデータを最大限に活用できるようにすることを1つの目標として、支援に全力を注いでいます。ネットアップは、信頼できるエンタープライズクラスのデータ サービスをクラウドにもたらし、またクラウドのシンプルな柔軟性をデータセンターにもたらします。業界をリードするネットアップのソリューションは、さまざまなお客様の環境や業界最大手のパブリック クラウドに対応します。

クラウド主導のData-Centricなソフトウェア企業であるネットアップは、お客様に最適なデータ ファブリックの構築をサポートし、クラウド対応をシンプルに実現し、必要なデータ、サービス、アプリケーションを適切なユーザにいつでも、どこからでもセキュアに提供できる唯一のベンダーです。

詳細については、**www.netapp.com/ja/**をご覧ください。