1 研究企画部門

LLM＋×IOWN
～NTT版LLMの誕生、IOWNの進展、そして2つの相互作用～

2023年11月、NTTは大規模言語モデル（LLM）「tsuzumi」を発表した。今後tsuzumiを活用した様々なサービス開発を進展させていく。また、2023年3月にAPNの本格サービスがスタートしたIOWNはさらなるフェーズに向かっている。本誌ではLLMの概要とIOWNの状況、IOWNとLLMによる相乗効果について紹介する。

日本電信電話株式会社
執行役員
研究開発マーケティング本部
研究企画部門長　**木下 真吾** 氏

――最近発表された大規模言語モデル（LLM）tsuzumiについて教えてください

今年11月1日にNTT独自の大規模言語モデル（LLM）であるtsuzumiを発表しました。tsuzumiの目指す方向性は何でも知っている巨大なLLMを目指すのではなく、専門知識をもった小さなLLMの集合を作ることです。NTT研究所は、日本語分野の自然言語処理研究においてはトップカンファレンス※採択数（2015年-2021年）で世界1位であり、その知見を活かし、我々はtsuzumiに、①モデルの軽量さ、②言語性能（特に日本語）の高さ、③柔軟なカスタマイズ、④マルチモーダル性という特徴をもたせました。①モデルの軽量性については、「軽量版」（70億パラメーター）と「超軽量版」（6億パラメーター）の2種類を発表しており、特に超軽量版はOpenAIのLLM「GPT-3」（175億パラメーター）と比較すると約300分の1にサイズを抑えています。モデルの軽量化は学習コストおよび推論コスト（ユーザーによるLLM利用コスト）を大幅に抑えられるほか、電力量削減にもつながります。例えばGPT-3規模の巨大なモデルの学習には約1300MWh、原発1基1時間分の電力量が必要と言われていますが、これを抑制することができます。②言語性能については、軽量版においてGPT-3.5の他、代表的な国産LLMの中でトップクラスの性能を示すことを確認しました。また英語の処理性能もMetaのLLM「Llama2」（70億パラメーター）と同程度の性能を実現しています。これらの特徴を生かし、ユーザーは電力量を抑えたサステナブルな環境でコストを抑えながら、ベースモデルの学習時点では対応していなかった業界の情報や最新情報にもチューニングによって柔軟に対応することができ（図1）、マルチモーダル性により自然言語だけでなく、文書や画像を提示しながらの質問にも応答できるようになります。

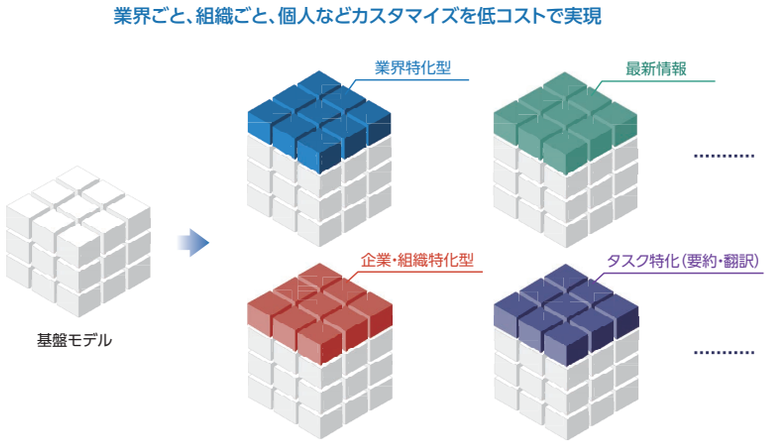

図1　tsuzumiの様々なチューニング

――IOWN構想とこれまでの進展について教えてください

NTTの掲げるIOWN構想は、光電融合デバイスを用いることによっ

て、低消費電力かつ大容量／高品質、低遅延のネットワーク基盤を実現しようとするものです。これまでに、まずAll-Photonic Network（APN）のプロダクト化の進展があり、IOWNの方式に準拠したAPN装置の販売がパートナー企業から開始されました。そしてこの装置を活用して、2023年3月にNTT東西がAPN IOWN 1.0としてネットワークサービスを提供開始しました。

これらを用いた遠隔のコンサート、Eスポーツ、お笑い、ダンス等、未来のエンタメと呼べるものの実証を行ってきましたが、このほか今後本格化していくターゲットに未来のデータセンターがあると考えています。現在、用地不足のためにデータセンターを都市部に新設することが難しくなっています。一方、郊外にデータセンターを増設する場合、主要都市間との遅延時間を考慮する必要があるため従来の方法では都心から約60km圏内にしかデータセンターを増設することができません。APNを活用することで、接続間距離を100kmに拡大することが可能となり、都心から少し離れた余った土地を活用してデータセンターを増設することができます。

――IOWN2.0と3.0の状況を教えてください

拠点のネットワーク装置間に留まらず、サーバ内のリソース間を直接APNで接続することで超省電力や高性能を達成するData Centric Infrastructure（DCI）の実現をめざしており、IOWN 2.0ではDCI step1として、遠隔のサーバ内のボード単位を光電融合デバイスで接続することによって超省電力・超高速なスイッチングを目指しています。このキーデバイスが光電融合デバイスの第3世代の光エンジンと呼ばれるものです。こちらは2025年度商用提供予定のスイッチボードで現在実用化に向けた試験を実施しております。IOWN 3.0ではさらにチップ単位を光電融合デバイスで接続するDCI step2となり、このキーデバイスとなる光電融合第4世代も2028年度の提供を予定しています。

――IOWNとLLMの相乗効果について教えてください

現在IOWNとLLMを組み合わせた実験を行っています。APNとLLMの組み合わせでは、実際に横須賀にある学習データを、100km離れた三鷹に配置したGPUクラウドからAPNによりリモートアクセスする環境を構築しました。インターネット接続であれば、LLMの学習が非常に遅くなってしまいますが、APN接続によりtsuzumiの学習は0.5%程度の性能低下で仕上げることができました。また、IOWN 2.0以降のコンピューティングにおいては、光スイッチを用いて光にダイレクトに繋ぐことにより、LLMの学習や推論プロセスをさらに効率化できる可能性があります。DCI step1/2との組み合わせによりGPUが使用されていない時間をなるべく減らして最小限の計算機リソースで実現することを狙っています。さらに将来の姿として、IOWNを活用したNTTの目指すAIの未来としてAIコンステレーションの実現を考えています（図2）。これは何でも知っている1つの巨大なLLMを作るのではなく、小さく専門性や個性を持ったLLMを複数組み合わせることによって、より賢く効率的に社会課題を解決するアーキテクチャとなります。

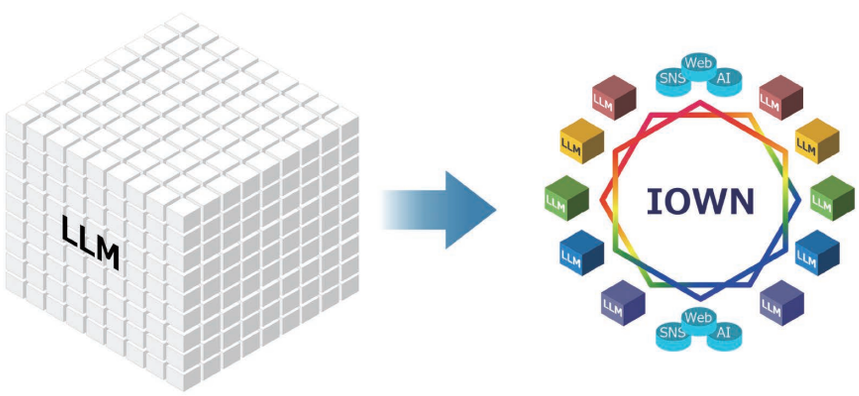

図2　AIコンステレーションの概念図

――最後に読者の方へ一言お願いします

このたび研究企画部門は、マーケティング部門、アライアンス部門と連携し、研究開発マーケティング本部として、2023年6月に体制を一新しました。NTTの研究開発は今後、世界最高峰の研究地位を確固たるものにするとともに、新体制のもと、顧客やパートナー企業の皆様の声をいち早く研究開発に反映し、IOWNやLLMなどの成果を世の中へ着実に社会実装する取り組みを加速していきたいと思います。

※TACL, NAACL, ACL, EMNLP, COLING