総務省・NICTが整備する学習用言語データのアクセス提供について

資料3－4

1．学習用言語データの整備について（NICTの現状/今後の方向性）

【NICTにおける学習用言語データの現状】

・ Webクローリング※による収集＋クレンジング　350GB

※公開中のWebサイトに機械的にアクセスし、情報を収集する手法（多くのAI開発者が活用）

・「学術研究目的」での収集　→　他社提供に当たっては、「共同研究」という形で提供を検討

【今後の方向性】（R6新規予算施策（３カ年計画）の目指すところ）

・ Webクローリングによる収集＋クレンジング　3.5TB（GPT-3相当、現状の10倍）

・ 当面は、NICTとの「学術研究目的の共同研究」という形での検討を進めつつ、今後、

① 共同研究機関が、開発成果（LLM）を用いた商用サービスを提供できるようにする

② 「共同研究」の形態に依らなくても学習用言語データを第三者へ提供できるようにする

ための方策について、法的な課題の整理を行う。（個人情報保護法、著作権法等）

2．アクセス提供対象の方向性

（１）学習用言語データを活用した犯罪巧妙化等の防止や、国内における開発力強化を図る観点から、学習用言語データへの適切なアクセスコントロールを実施。

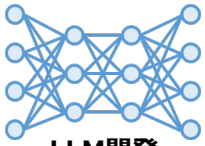
（２）国が実施する研究開発の運用等も参考としつつ、今後、具体的な内容について検討。

（例）・ AIガイドラインの遵守

・ 国内開発拠点の設置、国内での商用サービスの提供や責任ある体制の構築　等

（参考）AIの開発力強化に向けた取組

- NICTは、従来から多言語音声翻訳などAI自然言語処理に関する研究開発を実施してきたことから、AI学習に適した質の高い大量の言語データ構築に関する知見を有している。
- 総務省では、我が国で開発されるAIの安全、安心を確保するとともに、基盤的な開発力を国内に醸成するため、NICTが整備する学習用言語データを拡充し、民間企業やアカデミア等へのアクセスを提供するとともに、偽・誤情報への対応等を実施（令和６年度概算要求施策・10億円）。

	 **(1) LLM開発に必要となる学習用言語データの整備・拡充**	 **計算資源整備**	 **LLM開発**	 **(2) 偽・誤情報への対応等（周辺技術との連携等によるリスク対応）**
現状	容易に使用可能な学習用言語データは、米国団体が収集した公開データ等に依存。（日本語の量、品質、安全性に課題あり。）	民間企業による計算資源の整備の補助、公的機関の計算資源の拡充等により、国内の計算資源の整備・拡充が進展。	我が国の一部民間企業がLLM開発を実施中。	LLM等生成AIに起因する様々なリスクが指摘されているが、生成AI開発が先行し、リスクに対応するための技術の開発や社会実装は後手を取っている状況。
実施施策	① 日本語を中心とする**学習用言語データの大規模整備・拡充** ② **高品質で安全性の高い学習用言語データを作成**するための**データクリーニング技術**の開発 ③ **学習用言語データの高品質化**を図るための**LLMを用いた試行学習** ④ **学習用言語データへのアクセス提供枠組みの構築**	国際競争力を有するLLMの開発用の大規模計算資源の整備は関係省庁と連携して対応	国際競争力を有するLLMの開発は、民間企業等による取組を学習用言語データ整備によって支援	① ディープフェイク等、**悪意のある者による偽・誤情報の拡散等への対策技術**等の開発・実証 ② **LLMから出力された文章の信頼度の判定を可能とする技術**の開発・実証 ③ **LLMから出力された文章による著作権侵害の可能性を検知する技術**の開発・実証 等