形態素解析・係り受け解析AIにおけるデータ管理とデモ環境の統合

安岡孝一 (京都大学人文科学研究所附属人文情報学創新センター)

BERT / RoBERTa / DeBERTa / GPT-2 等の事前学習モデルを用いた形態素解析・係り受け解析エンジンは、大量のテキストデータと、アノテーションデータを必要とする。テキストデータから mdx で事前学習モデルを構築し、Universal Dependencies (UD) にもとづくアノテーションデータでファインチューニングをおこない、Jupyter (Google Colaboratory) 上にデモ環境を構築する、というのが、われわれがおこなっている作業手順である。さらに、このデモ環境 (UD エディター) を使って、さらなるアノテーションデータを作成し、解析エンジンとデモ環境をどんどん更新していって、解析精度を上げていく、というのが、われわれの AI データエコシステムである。

すなわち、われわれが用いるテキストデータ・アノテーションデータ・事前学習モデル・解析エンジン・デモ環境は、常に更新されている。どのデータからどのエンジンを構築したのかバージョン管理すべく、われわれは全てを Git リポジトリに記録し、専用の GitLab サーバーを人文情報学創新センターで運用している。また、Google Colaboratory でデモ環境を動作させるため、Jupyter ノートブックと周辺プログラムは GitHub に、事前学習モデルと解析エンジンは HuggingFaceHub に、それぞれ Git ブランチを置いている。

本事業は 2023 年度いっぱいで終了した (継続申請が通らなかった) が、2025 年 2 月以降、継続の危機に瀕している。HuggingFaceHub が運用方針を変更し、われわれの解析エンジンを、しばしば「unsafe」とみなすようになった点が大きい。

2025 年 1 月 30 日リリースの pytorch 2.6 では、`torch.load` のデフォルトが `weights_only=False` から `weights_only=True` に変更された。セキュリティがらみの変更だが、われわれの解析エンジンは、`torch.load` 時に pickle 形式のプログラムを読み込んでいるため、`weights_only=False` の必要がある。そうしたところ、Protect AI が HuggingFaceHub において、われわれの解析エンジンを「unsafe」と指摘し始めたのである。不名誉きわまりない。

われわれ自身も検討を加えたが、われわれの解析エンジンには CVE-2025-32434 のセキュリティホールは存在せず、「unsafe」よばわりは不適切である。その旨を Protect AI に伝えたところ、いくつかの解析エンジンは「safe」に戻ったが、いくつかの解析エンジンは同一構造にもかかわらず「unsafe」のままだった。理由の説明はない。仕方ないので、HuggingFaceHub で「unsafe」となった解析エンジンを「改造」し、プログラムを外部に追い出すことで `weights_only=True` としたところ、Protect AI も「safe」とみなした。しかし、この「改造」で必要となった計算量と電気代は、われわれの持ち出しとなった。

非常に残念だが、これが本事業の顛末である。われわれは、他の資金を調達することで、新たな事前学習モデルと解析エンジンを構築中であり、そこでは、本事業で「unsafe」となった解析エンジンでの経験が、新たな解析エンジンの設計に活かされている。しかしながら、その資金そのものを、本事業で製作した解析エンジンの「改造」に充てるのは、やはりどう考えてもマズイ気がする。本事業で製作した解析エンジンが「unsafe」だというなら、資金が無い以上、その解析エンジンを迷わず削除するのが、AI セキュリティにおける「正しい」態度なのだろう。しかし、それは本事業の終了を意味する。何とも難しいところである。