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、われわれが現在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。

この GitLab サーバーを京都大学の外に置いた場合、著作権の問題が起こりうる。GitLab サーバーは「公衆送信」をおこなっており、著作権法第 23 条の対象である。たとえばアイヌ語では、金田一京助 (1882–1971) や久保寺逸彦 (1902–1971) の著作権が 20 年延長されたため、彼らが解析したアイヌ語テキストは、現在も第 23 条の保護下にある。もちろん第 30 条の 4・第 47 条の 4 により、mdx で「利用」するのは OK である。しかし「公衆送信」をおこなうには、われわれの判断では、第 35 条「授業」のために「公衆送信」するしかなく、大学内に GitLab サーバーを設置している。

では、この Git 環境を、Gakunin RDM 配下に置くことが可能だろうか。GitLab も GitHub も HuggingFaceHub も、それぞれ API トークンによるアクセスを可能としている。この API トークンを、Gakunin RDM の eduPersonPrincipalName から発行すれば、Gakunin RDM から各 Git 環境へのアクセスが可能となる。しかし、その手法は、技術的には実現可能であるものの、実運用に耐えなかった。Gakunin RDM は「メンバーの追加」や「メンバーの削除」には対応しているが、「メンバーの異動」を想定していないのだ。

アノテーションデータの開発は、かなり長期間に渡ることから、期間中にメンバーの異動がしばしば起こる。たとえば 2023 年 4 月に、京都大学から国文学研究資料館へ異動したメンバーを考えてみよう。京都大学の IdP は、もちろん使えなくなった。しかし、国文学研究資料館は、この時点では Gakunin RDM に参加しておらず、Gakunin RDM 経由でのアクセスは閉ざされた。2023 年 7 月 27 日に国文学研究資料館は Gakunin RDM に参加したが、そこで降りてくる eduPersonPrincipalName は、もちろん京都大学とは異なっており、ここの連続性を保とうとすると、GitLab サーバー側で「かなり汚い」設定をおこなうことになる。ならば Gakunin RDM など使わず、最初から最後まで GitLab サーバー上の Username を使い続ける方がむしろ安全、というのが、われわれの現時点での結論である。