

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東アジア人文情報学研究センター)

BERT / RoBERTa / DeBERTa 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く、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 サーバーを東アジア人文情報学研究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。

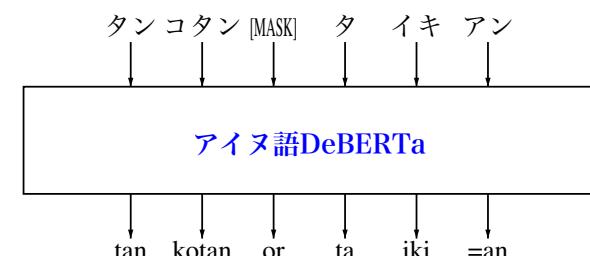
では、この環境を Gakunin RDM 配下に置くことが可能だろうか。事前学習モデル構築とファインチューニングは mdx 上なので、Gakunin RDM 配下だと考えていいだろう。デモ環境構築については、Gakunin RDM のデータ解析機能に Jupyter があるので、「30 日間はとおくと消える」点に注意すれば大丈夫そうだ。GitLab サーバーは、mdx 上にもインストール可能なので、API トークンを Gakunin RDM の eduPersonPrincipalName から発行するやり方で、何とかなる。何とかなると思ったが、甘かった。Gakunin RDM は「メンバーの追加」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。

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。でも、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、ということになってしまう。

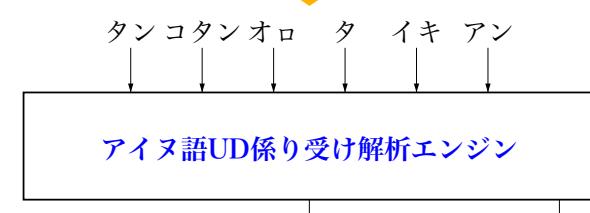
また、GitLab サーバーを京都大学の外に置いた場合、著作権の問題が起りうる。GitLab サーバーは「公衆送信」をおこなっており、著作権法第 23 条の対象である。たとえばアイヌ語では、金田一京助 (1882–1971) や久保寺逸彦 (1902–1971) の著作権が 25 年延長されたため、彼らが解析したアイヌ語テキストは、現在も第 23 条の保護下にある。もちろん第 30 条の 4・第 47 条の 4 により、mdx で「利用」するのは OK である。しかし「公衆送信」をおこなうには、現在のわれわれの判断では、第 35 条「授業」のために「公衆送信」するしかなく、とりあえず大学内に GitLab サーバーを設置した。ただし、Gakunin RDM が第 31 条の適用を受けてくれるのなら、それはそれで方策があると思われる。

テキストデータ	
- 2 -	- 3 -
<p>Kamuchiakip kamui yaieyukara, "Shirokamipe ranran piishan."</p> <p>"Shirokamipe ranran piishan, konkai- ranan piishan" aran seko chiki kame- peteso saqud aina, amuketan entasafue chikishi kor shichopokan intaneshi ko- teeta wekuri tanu nishuu ne, teeta nisgu mashu amuketan entasafue chikishi kor. Atsukessan ta amuketan entasafue chikishi kor. "Shirokamipe ranran piishan, konkai- ranan piishan" aran seko chiki kame, hechuchitar entasafue chikishi kor, amuketan entasafue chikishi kor. "Perla chikappo" kamui chikappo!</p> <p>Kels betata, alashad wa ton chikappo kamui chikappo tulan wa antuk, loschukkar sono rametok shiro chikappo ne ruwe tapan" hawoloi kane, teeta wotukun tanu nishuu nep poutari, konkai ponku konkai ponku.</p> <p>(1) 皆は男を多く多くなる、みんな少しあがつて喰へます。 手扱はそれで樹木や鳥などに前に歯で齧る。知らず少子の申し</p>	<p>翁の神の自ら歌った詩 銀の鏡の鏡の歌うまほりに</p> <p>「鏡の鏡はまほりに、翁の鏡 歌る歌はまほりに、さる云ひ翁は我ひながら 流に流つて下り、人間の歌の上を 通しながら下を歌めるさる 翁の歌の歌はまほりに、さる云ひ翁は我ひながら 今も歌は人間の歌をかきもやしの歌に 歌ひながら下を歌めるまほりにであります。 「鏡の鏡はまほりに、翁の鏡 歌の歌はまほりに、さる云ひ翁は我ひながら 歌ひながら下を歌めるさるに、さる云ひ翁を 歌ひながら下を歌めるさるに、さる云ひ翁は我ひながら 通すまほりに、さる云ひ翁は我ひながら 云ふまほりに、翁の鏡の歌はまほりに</p> <p>「翁の鏡はまほりに、翁の鏡の歌はまほりに さる、翁を歌ひ翁の歌 神様の歌を歌でたものは、一ぱんさきに歌つた者は ほんとうの勇者はほんとうの歌者だ」 云ひながら、音質入出で歌が持てになつてゐる者の 子等は、歌ひながら下を歌めるまほりに</p> <p>「翁の鏡はまほりに、翁の鏡の歌はまほりに さる、翁を歌ひ翁の歌 神様の歌を歌でたものは、一ぱんさきに歌つた者は ほんとうの勇者はほんとうの歌者だ」 云ひながら、音質入出で歌が持てになつてゐる者の 子等は、歌ひながら下を歌めるまほりに</p>

mdxで事前学習モデル構築

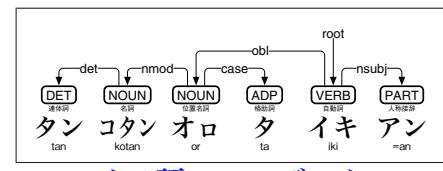


mdxでファインチューニング



root				
det	root	conj	case	
det	nmod	root	case	
			root	
nssubj	obl		root	nssubj
nssubj	obl		advcl	root

Jupyterによるデモ環境構築



アイヌ語 UD エディター