報告

教育データを大規模言語モデルに学習させるための社会的環境構築調査

Investigation of Building a Social Environment for Learning Data into a Large-scale Language Model

藤井 政登 FUJII Masato
デジタルハリウッド大学大学院 院生
Digital Hollywood University, Graduate School

佐藤 昌宏 SATO Masahiro
デジタルハリウッド大学 教授 学長補佐
Digital Hollywood University, Professor, Advisor to the President

学びの個別最適化のためには、教育データの利用が必要になる。また、大規模言語モデルの普及に伴い、言語モデルの作成に必要なデータセットの作成が始まっている。しかし、教育データのモデル構築のための試みは未だなされていない。教育データは、スキルセット作成の前提であり、その整備は海外では進んでいるが、国内では社会的な環境に問題を抱えている。この報告では、スキルセットデータを作成するために必要な社会的な環境整備を扱う。個人情報である、教育データと教育用IDについて海外・国内での整備状況の調査を行った。また、スキルセットを構築する際のユースケース検討も併せて継続している。2023年1月に提出した修士論文作成後の継続調査の報告を行う。

1. はじめに

1.1 大規模言語モデルに教育データは含まれていない

近年、ChatGPTなどの大規模言語モデルが社会に浸透してきている。その利用方法については各所で解説や議論がされているが、言語モデル構築のためのデータセットについての議論はあまりされていない。

そんな中でも、オープンソースの言語モデルの作成やデータ公開が続いている。「The Pile」[1]や、Metaの「Llama 2」[2]、東京大学が公開した「Weblab-10B」[3]など、オープンソースのデータセットの公開が続いている。しかし、データセットが主としてインターネット上のテキストデータから構成されているため、教育データは含まれていない。中でも、スキルセットのデータ構築は遅れている分野といえる。

1.2 スキルセットデータとは何か

スキルセットとは、個人が持っている技能や能力の総体を指す言葉であり、特定の職務やタスクを遂行するための技術や知識、経験などの組み合わせを指す。プログラミング言語、機械操作、外国語などの具体的なスキルであるハードスキル、人間関係やコミュニケーション、問題解決能力、リーダーシップ、時間管理などを扱うソフトスキルの2つに大きく分類できる。これをデータ化する場合、どのようなデータが対象になるかユースケースを集め5つに分類した。

(1) 技術スキル：プログラミング言語、デザインツール、機械操作など、特定の職務で必要とされる具体的な技術
(2) ソフトスキル：コミュニケーション、問題解決、リーダーシップなどの人間関係や対人スキル
(3) 専門知識：特定の業界や分野における深い知識や経験。例：金融、法律、医療など
(4) 認証や資格：必要な資格や認証、許可証など
(5) 経験：過去の職歴やプロジェクト経験など

分類後、これらのスキルセットが利用できるシーンを教育の観点からユースケースを集め分類した。

(1) 採用：企業が求職者のスキルセットと必要な職務の要件を照会する際
(2) キャリア開発：個人が自身のキャリアパスを計画する際に、必要なスキルセットを把握するため
(3) チーム編成：プロジェクトの成功のために、異なるスキルセットを持つメンバーを組み合わせる際
(4) 教育・トレーニング：個人や組織が必要なスキルセットのギャップを特定し、教育プログラムを計画する際

教育の個人最適化を社会性と両立するためには、このようなスキルセットデータの利用が必須であり、またそれを社会に浸透させるためには、大規模言語モデル構築のためのスキルデータの作成が急務である。

1.3 スキルセット作成に必要な個人識別ID

スキルセットデータを構築するためには、ある個人が学習した学習履歴と、個人の持つとされるスキルや、組織の中でのポジションなどのデータを紐づけてモデルを作る必要がある。この際に個人のデータを扱うための個人識別IDが必要であり、教育データ活用にも必ず必要になる。

2023年3月から、教育用の個人識別IDの調査を開始した。範囲は、海外と国内両面にわたり、継続的な調査を行っている。

2. 既存の教育ID調査

2.1 海外での教育ID利用調査

教育分野では、個人識別IDを、SIS（Student Information System）と呼ぶことが多い。2023年3月にインターネットを使った海外でのSISの利用状況を調査した[4]。IDのあるなし、IDが国民IDと同じなのか、独自の教育用IDを持っているか、LMS（Learning Management System）と同じIDでの利用が可能かどうか、他のシステムとのデータ互換性などを考慮に入れているかどうかを「標準化対応」と考え、まとめたものが表1である。インドネシアを除きほぼ全ての国が、教育用のIDを持ち、運営していることがわかる。

表1：IDシステムを持っていないのは、調査国の内、日本とインドネシアのみ

項番	国名	データの収集範囲	ID付与	校務データ収集	学習ポータル	標準化対応
1	米国	州単位	教育ID	○	○	○
2	フィンランド	国全体	国民ID	○	○	○
3	オランダ	国全体	国民ID	○	○	○
4	デンマーク	国全体	国民ID	○	○	○
5	フランス	国全体	国民ID△	○	✗	△
6	英国	国全体	教育ID	○	✗	○
7	オーストラリア	州単位	教育ID	○	✗	○
8	ニュージーランド	国全体	教育ID	△	✗	✗
9	シンガポール	国全体	教育ID	○	○	○
10	インドネシア	✗	✗	✗	✗	✗
11	中国	年単位	✗	✗	○	✗

2.2 米国におけるSISシステム

米国では、「全米共通学力基準」(Common Core) を制定。2010年から全米で数学と読解力の共通テストを行っている。NCLB法のもと、全米の初等・中等教育課程 (6歳～18歳までの12年間) の公立学校のデータを集めるシステムとしてEDFactsが2004年に稼働を開始。SISから州のシステム「SLDS (Statewide Longitudinal Data Systems)」経由で生徒データを収集し、2018年度の時点では1万8,617学区、5,069万人の生徒の統計データが収蔵されている。また、一部はオープンデータとして学校ごとの人種別のドロップアウト率や、算数テストの得点などのデータが公開されている[5]。

2.3 国内教育ID利用と学習指導要領コード

個人識別ID、教育用IDについては、国内には存在しない。インドネシアと同じく高等教育機関や、一部の私学で、民間の校務データシステムや、オープンソースのシステムが運用されている。単位の互換などのシステムもない。ただ、校務システムに関するガイドラインのようなものは文部科学省から出されている[6]。しかし、利用は学校単位で、学外を対象としたデータ連携を行えるものではない。

しかし、学習指導要領コードだけは制定された[7]。範囲は、幼・小・中・高・特別支援の全ての教科である。このコードにより、何を学習したのかをコードとして扱えるようになった。しかし、このデータを個人に紐づけるシステムには、教育用IDの制定が必要である。

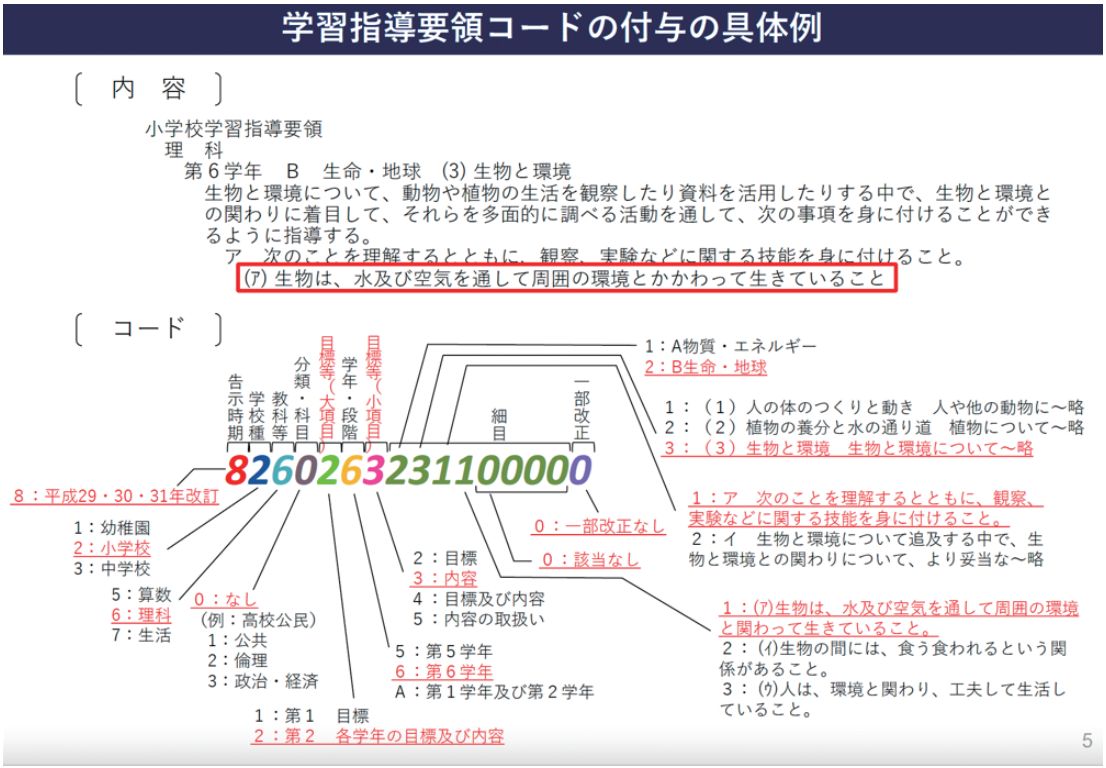

図1：指導内容ごとにコードが振られ学習内容の把握が可能に[7]

3. 大規模言語モデルを使ったモデル構築

3.1 Transformerとスキルセット

ChatGPTやGoogleのBardに使われているTransformerは、単語間のベクトル距離をAttention Mechanismを使って評価する。InputとMemoryが同一のAttentionとして評価されるため、スキルセットのような曖昧な概念に結びつく単語感の距離評価に最適である。

従来のエキスパートシステムよりも一般化されたスキルセット。例えば、「英語が得意」といった曖昧な表現も扱うことができ、モデルを構築できる。

IDにより取得された個人のデータ、学習履歴、成績評価などに加え、テキストで評価されたスキルなども大規模言語モデルのデータとしてモデルを構築できるのではないかと考えている。

4. おわりに

Transformerアーキテクチャをベースにした新しいモデルをゼロからトレーニングするには、大量の計算リソースとデータが必要である。しかし、オープンソースのデータセットや、それに関連したライブラリー提供も始まっているため、個人でも数年以内には、モデルの構築が可能になるのではないかと考えている。

そのため、GPUや大規模言語モデルの学習を集中的に始めている。しかし、その前提には教育データが大量にあることが必要であり、そうした社会構築が進むことを願っている。

参考文献

[1] The Pile is a 825 GiB diverse, open source language modelling data set that consists of 22 smaller, high-quality datasets combined together.
https://pile.eleuther.ai/ (Accessed 2023-08-22)
[2] Meta、新たな大規模言語モデル「Llama 2」 商用利用可でGPT-3.5に匹敵
https://www.watch.impress.co.jp/docs/news/1517245.html (Accessed 2023-08-22)
[3] 100億パラメータサイズ・日英2ヶ国語対応の大規模言語モデル“Weblab-10B”をオープンソースで公開しました。
https://weblab.t.u-tokyo.ac.jp/100%E5%84%84%E3%83%91%E3%83%A9%E3%83%A1%E3%83%BC%E3%82%BF%E3%82%B5%E3%82%A4%E3%82%BA%E3%83%BB%E6%97%A5%E8%8B%B12%E3%83%B6%E5%9B%BD%E8%AA%9E%E5%AF%BE%E5%BF%9C%E3%81%AE%E5%A4%A7%E8%A6%8F%E6%A8%A1/ (Accessed 2023-08-22)
[4] Danish Center for Big Data Analytics driven Innovation: DABAI
https://www.dabai.dk/
Kennisnet
https://www.kennisnet.nl/nummervoorziening/
Centre for the Science of Learning & Technology: SLATE
https://www.uib.no/en/slate
Avain Parempaan Oppimiseen Ammattikorkeakouluissa
https://esignals.fi/kategoria/digitaalisuus/avain-parempaan-oppimiseenammattikorkeakouluissa/#190cf42e
Verein Forum Neue Medien in der Lehre Austria https://fnma.at/
Spanish Network Of Learning Analytics: SNOLA https://snola.es/
Kungliga Tekniska högskolan: KTH https://skr.se/skr.25.html
the METAL project
https://capture.dropbox.com/ezzsH9HjR1LY0FdY
インドネシア国 高等教育・職業教育にかかる情報収集・確認調査最終報告書(2022年1月独立行政法人国際協力機構(JICA))
https://openjicareport.jica.go.jp/247/247/247_108_12367314.html
National Student Index(NSI)
https://capture.dropbox.com/PtU7Hrk6JMAmMNht
OneSchool
https://capture.dropbox.com/UcJzmGsi4jHdJm3I
EEF(Education Endowment Foundation) https://educationendowmentfoundation.org.uk/
令和元年度学びと社会の連携促進事業(学習ログ等の活用に向けた収集すべき標準項目等の素案の作成等)経済産業省教育産業室最終報告書
https://www.meti.go.jp/meti_lib/report/2019FY/000191.pdf
What Works Clearinghouse(WWC)
https://capture.dropbox.com/u7J8HhKSzXT9PaiY
以上、いずれも(Accessed 2023-03-27)
[5] EDFacts
https://www2.ed.gov/about/inits/ed/edfacts/index.html (Accessed 2023-08-22)
[6] 文部科学省 教育の質の向上に向けた効果的なデータ連携・活用のポイントと学校改善事例集
https://www.mext.go.jp/content/1387543_02.pdf
[7] 学習指導要領コードについて 初等中等教育局 学びの先端技術活用推進室 6頁
https://www.mext.go.jp/content/20201016-mxt_syoto01-000010374_3.pdf