EdNet: A Large-Scale Hierarchical Dataset in Education

Youngduck Choi\textsuperscript{1,2}, Youngnam Lee\textsuperscript{1}, Dongmin Shin\textsuperscript{1}, Junghyun Cho\textsuperscript{1}, Seoyon Park\textsuperscript{1}, Seewoo Lee\textsuperscript{1,3}, Jineon Baek\textsuperscript{1,4}, Chan Bae\textsuperscript{1,3}, Byungsoo Kim\textsuperscript{1(2)}, and Jaewe Heo\textsuperscript{1}

\textsuperscript{1} Riiid! AI Research, Seoul, Republic of Korea
\textsuperscript{2} Yale University, New Haven, USA
\textsuperscript{3} UC Berkeley, Berkeley, USA
\textsuperscript{4} University of Michigan, Ann Arbor, USA

Abstract. Advances in Artificial Intelligence in Education (AIEd) and the ever-growing scale of Interactive Educational Systems (IESs) have led to the rise of data-driven approaches for knowledge tracing and learning path recommendation. Unfortunately, collecting student interaction data is challenging and costly. As a result, there is no public large-scale benchmark dataset reflecting the wide variety of student behaviors observed in modern IESs. Although several datasets, such as ASSISTments, Junyi Academy, Synthetic and STATICS are publicly available and widely used, they are not large enough to leverage the full potential of state-of-the-art data-driven models. Furthermore, the recorded behavior is limited to question-solving activities. To this end, we introduce EdNet, a large-scale hierarchical dataset of diverse student activities collected by Santa, a multi-platform self-study solution equipped with an artificial intelligence tutoring system. EdNet contains 131,417,236 interactions from 784,309 students collected over more than 2 years, making it the largest public IES dataset released to date. Unlike existing datasets, EdNet records a wide variety of student actions ranging from question-solving to lecture consumption to item purchasing. Also, EdNet has a hierarchical structure which divides the student actions into 4 different levels of abstractions. The features of EdNet are domain-agnostic, allowing EdNet to be easily extended to different domains. The dataset is publicly released for research purposes. We plan to host challenges in multiple AIEd tasks with EdNet to provide a common ground for the fair comparison between different state-of-the-art models and to encourage the development of practical and effective methods.

Keywords: Dataset · Education · Artificial intelligence · AIEd · Knowledge tracing

© Springer Nature Switzerland AG 2020
I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 69–73, 2020.
https://doi.org/10.1007/978-3-030-52240-7_13
Fig. 1. A possible scenario of a student using Santa and example student data in EdNet. After the student purchases a 50-day pass (p25), they solve an LC question (q878). The timestamps at which they played and paused audio were recorded. They also eliminated ‘a’ and chose ‘c’ as an answer.

1 Introduction

In this paper, we introduce EdNet\textsuperscript{1}, a large-scale hierarchical dataset consisting of student interaction logs collected over more than 2 years from Santa\textsuperscript{2}, a multi-platform, self-study solution equipped with artificial intelligence tutoring system that aids students in preparing for the TOEIC® (Test of English for International Communication) test. To the best of our knowledge, EdNet is the largest dataset open to the public, containing 131,441,538 interactions from 784,309 students. Aside from question-solving logs, EdNet also contains diverse student behaviors including but not limited to self-study activities, choice elimination, and course payment. EdNet has a hierarchical structure where the possible student actions in Santa are divided into 4 different levels of abstraction. This allows the researcher to select the level appropriate for the AIEd task at hand, for example, knowledge tracing or learning path recommendation.

\textsuperscript{1} https://github.com/riiid/ednet.

\textsuperscript{2} https://santatoeic.com.
Table 1. Comparison of Datasets in Education (ASSISTments [2, 4], Synthetic-5 [5], Junyi Academy [1], STATICS-2011 [3] and EdNet). Here *logs* stands for interactions, and Synthetic-5, Junyi Academy, and Statics-2011 are renamed as Syn-5, Junyi, and Stat-2011.

|          | ASSISTments | Syn-5 | Junyi | Stat-2011 | EdNet |
|----------|-------------|-------|-------|-----------|-------|
|          | 2009        | 2012  | 2015  | KT1       | KT2   |
| # of     | 4,217       | 46,674| 19,917| 4,000     | 247,606|
| students |             |       |       | 335       | 784,309|
| # of     | 26,688      | 179,999| 100  | 50        | 722   |
| questions|             |       |       | 1,362     | 13,169|
| # of     | 123         | 265   | –     | 5         | 41    |
| tags     |             |       |       | 27        | 188   |
| # of     | 0           | 0     | 0     | 0         | 0     |
| lectures |             |       |       | 0         | 0     |
| # of logs| 346,860     | 60,123| 2,708 | 73,612    | 361,092|
| types of | 3           | 3     | 1     | 1         | 5     |
| logs     |             |       |       | 3         | 4     |
| Public   | Yes         | Yes   | Yes   | Yes       | Yes   |
| available|             |       |       | No        | Yes   |
| Contents | No          | No    | No    | No        | No    |
| available|             |       |       | Yes       | Yes   |
| From     | Yes         | Yes   | Yes   | Yes       | Yes   |
| real-world|           |       |       | Yes       | Yes   |
| Collecting| 1y         | 1y    | 1y    | –         | 2y 2m|
| period   |             |       |       | 4m        | 2y 7m|
|          |             |       |       | 1y 3m     | 1y 3m|
|          |             |       |       | 1y 3m     | 1y 3m|

2 EdNet

*EdNet* is a dataset consisting of all student-system interactions collected over a period spanning two years by *Santa*, a multi-platform AI tutoring service with approximately 780,000 students in South Korea. *Santa* is available through Android, iOS and the Web. It aims to prepare students for the TOEIC (Test of English for International Communication®) Listening and Reading Test. Each student communicates their needs and actions through *Santa*, to which the system responds by providing video lectures, assessing their response or giving expert commentary. *Santa*’s UI and data-gathering process is described in Fig. 1. As shown in the figure, the *EdNet* dataset contains various features of student actions such as the identity of the learning material consumed or the time spent by the student in solving a given problem. The following subsections describe properties of *EdNet*.

2.1 Large-Scale

*EdNet* is composed of a total of 131,441,538 interactions collected from 784,309 students of *Santa* since 2017. Each student has generated an average of 441.20 interactions while using *Santa*. Based on those interactions, *EdNet* makes it possible for researchers to access large-scale real-world IES data. Moreover, *Santa* provides a total 13,169 problems and 1,021 lectures tagged with 293 types of

---

3 More detailed description of *EdNet* can be found in [https://arxiv.org/abs/1912.03072](https://arxiv.org/abs/1912.03072).
skills, and each of them has been consumed 95,294,926 times and 601,805 times, respectively. To the best of our knowledge, this is the largest dataset in education available to the public in terms of the total number of students, interactions, and interaction types (Table 1).

![Fig. 2. Hierarchical structure of EdNet.](image)

### 2.2 Diversity

*EdNet* offers the most diverse set of interactions among all existing public IES datasets (Table 1). The set of behaviors directly related to learning is also richer in *EdNet* than in other datasets, as *EdNet* includes learning activities such as reading explanations and watching lectures which aren’t provided in other datasets. The richness of the data enables researchers to analyze students from various perspectives. For example, purchasing logs may help analyze student’s engagement with the learning process.

### 2.3 Hierarchy

*EdNet* is organized into a hierarchical structure where each level contains different types of data points as shown in Fig. 2. To provide the various types of data in a consistent and organized manner, *EdNet* offers the data in four different datasets named KT1, KT2, KT3, KT4. As the postfix index of the datasets increases, the number of actions and types of actions involved also increase as shown in Table 1.

### 2.4 Multi-platform

In an age dominated by various devices spanning from personal computers to smartphones and AI speakers, IESs must offer access from multiple platforms in order to stay competitive. Accordingly, *Santa* is a multi-platform system available on iOS, Android and Web and *EdNet* contains data points gathered from both mobile and desktop users. *EdNet*’s platform-agnostic design allows the study of AIEd models suited for future multi-platform IESs, utilizing the data collected from different platforms in a consistent manner.
Acknowledgement. The authors would like to thank all the members of Riiid! for leading the Santa service successfully. EdNet could not have been compiled without their efforts.

References

1. Chang, H.S., Hsu, H.J., Chen, K.T.: Modeling exercise relationships in e-learning: a unified approach. In: EDM, pp. 532–535 (2015)
2. Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User Model. User-Adap. Interact. 19(3), 243–266 (2009)
3. Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the EDM community: the PSLC DataShop. In: Handbook of Educational Data Mining, vol. 43, pp. 43–56 (2010)
4. Pardos, Z.A., Baker, R.S., San Pedro, M.O., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. J. Learn. Anal. 1(1), 107–128 (2014)
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in Neural Information Processing Systems, pp. 505–513 (2015)