Yang, J. C. et al. (Eds.) (2018). Proceedings of the 26th International Conference on Computers in Education. Philippines: Asia-Pacific Society for Computers in Education

SCROLL Dataset in the Context of Ubiquitous Language Learning

Hiroaki OGATA[a*], Kousuke MOURI[b], Noriko UOSAKI[c], Mohammad Nehal HASNINE[a], Victoria ABOU-KHALIL[a], & Brendan FLANAGAN[a]

[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Institute of Engineering, Tokyo University of Agriculture and Technology, Japan*
[c]*Center for International Education and Exchange, Osaka University, Japan*
*hiroaki.ogata@gmail.com

Abstract: This paper introduces SCROLL dataset, a dataset that consists of foreign language learners lifelong learning experiences (i.e. lifelogs). Logs are chronologically collected from a context-aware ubiquitous language learning system called SCROLL (System for Capturing and Reminding of Learning Logs). The dataset contains ubiquitous learning logs from the period of November 2009 to March 2018 (except the years 2015 and 2016). Furthermore, we propose some baseline approaches using the provided information from the dataset. With these baselines, we target at providing references for other approaches that intends to solve problems in the domains of educational technology. The dataset can be used for educational research purposes, particularly for knowledge discovery using learning analytics and educational data mining strategies.

Keywords: Educational Data Mining, Learning Analytics, SCROLL Dataset, Ubiquitous Learning Logs, Ubiquitous Language Learning Research

1. Introduction

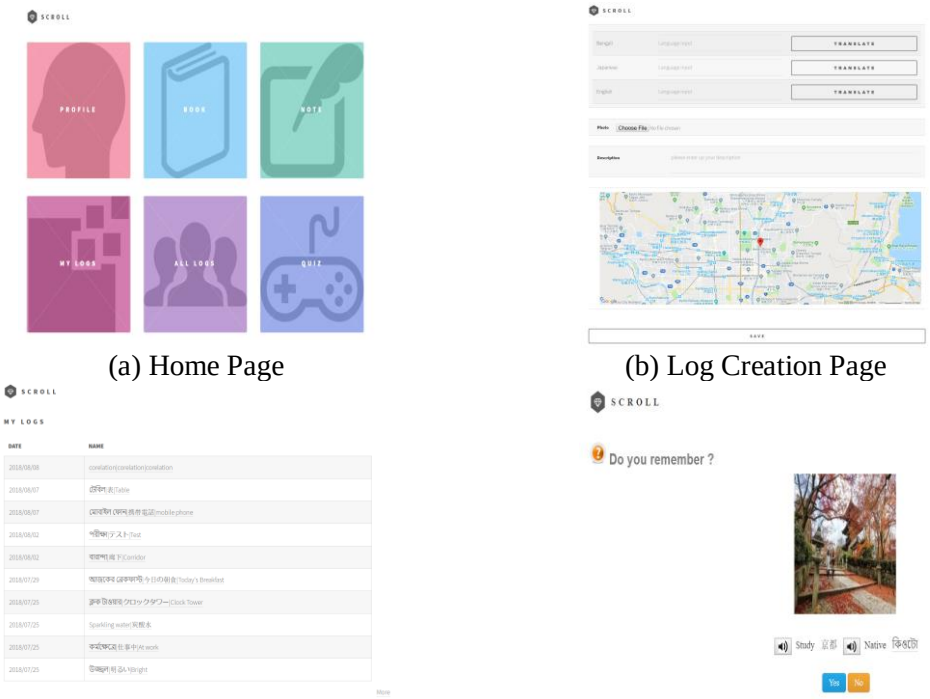
Lifelogging represents the phenomenon whereby learners can digitally capture their personal lives in varying amount of details, for a variety of purposes (Gurrin, Smeaton, & Doherty, 2014). With the availability of new technologies to track human's life activities, lifelogging (often addresses as personal big data) has become a mainstream research topic. Lifelogging may offer benefits to content-based information retrieval, retrieval of context, browsing, search, linking, summarization, and user interaction (Gurrin et al., 2014). Besides, lifelong data are used in various fields including healthcare, fashion, social network etc. In the language learning context, the concept of ubiquitous learning has been an important and breakthrough research topic. In recent years, ubiquitous computing is considered as the new hype in the information and communication that is normally associated with a large number of small electronic devices (small computers) which have computation and communication capabilities such as smart mobile phones, contactless smart cards, handheld terminals, sensor network nodes, radio frequency, identification (RFIDs) etc. which are being used in our daily life (Yahya, Ahmad, & Jalil, 2010) (K. Sakamura & N. Koshizuka, 2005). In general, ubiquitous learning uses the infrastructure of ubiquitous computing in a specific manner by which learning and teaching processes can be enhanced. Logs captured by ubiquitous technologies can be used for a wide range of applications including education, healthcare, security, surveillance, human-machine interaction, sports science etc. (Zhang & Sawchuk, 2012).

In our study, Ubiquitous Learning Log (ULL) is defined as a digital record of what learners have learned in the daily life using ubiquitous technologies (Ogata et al., 2011). The collection of ULLs was started under the project named "ubiquitous learning log (ULL)" in order to store intentionally what learners of foreign languages have learned in their daily life and consequently reuse them. In the collection of ULLs, two basic fundamentals of computer-supported ubiquitous learning, namely how to record and share learning experiences that happen anytime and in any place? and how to retrieve and reuse them for future learning? were taken into consideration (Ogata et al., 2014). In this paper, we describe a dataset consisting of ubiquitous learning logs that are collected from the perspectives of the foreign language learners. Those logs are accumulated by

using a context-awarded ubiquitous language learning system named SCROLL (System for Capturing and Reminding of Learning Logs). In this work, our primary focus is on developing a dataset for ubiquitous language learning research. The lack of large, publicly available and general purpose ubiquitous language learning datasets motivated us to build our own dataset. So far, several scientific researches in the domain of language learning with a partial use of this dataset have been carried out. As learning analytics, educational data mining and artificial intelligence mature, our aim is that this dataset will play a significant role in facilitating research in other related scientific domains.

2. Data Source

Ubiquitous Learning Logs (ULLs), described in this paper are collected chronologically from a context-aware ubiquitous language learning system called SCROLL (Ogata et al., 2011). SCROLL system is proposed to overcome one of the challenges of Computer Supported Ubiquities Learning (CSUL) research, namely capturing what learners have learned with the contextual data, and reminding the learners of it in the right place and the right time. ULLs are defined as the record of what learners have learned in their daily life using ubiquitous technologies (Ogata et al., 2011). Learners using SCROLL system are allowed to log their authentic learning experiences with photos, audios, videos, location, QR-code, RFID tag and sensor data, and afterward share them with other learners (Ogata et al., 2011). Also, learners are allowed to reuse those logs. SCROLL system is designed in a way that meaningful knowledge from past learning experiences can be extracted so that it can be used as a guide for learners' future behaviors. For this, LORE (Log-Organize-Recall-Evaluate) model was proposed which worked as the backbone of the system (Ogata et al., 2011). The SCROLL system runs on different platforms including personal computer, tablets, mobile devices and Android phones. Interfaces of the SCROLL system is shown in Figure 1.



(a) Home Page (b) Log Creation Page

(c) My Log Page (d) Quiz Page

Figure 1: Interfaces if the SCROLL System

3. The Dataset

This dataset, in the context of ubiquitous learning, consists of 27507 logs that are collected using SCROLL system from the period of November 2009 to March 2018 (except the years 2015 and 2016). The system has been used by users from over 30 different nationalities. However, in this paper, we report about the logs that are collected from learners registered themselves as Chinese,

English and Japanese languages as their default languages. Table 1 shows the summary of the dataset.

Table 1: Summary of the Dataset

Native Language	Number of Users	Total Logs
Chinese	104	3279
English	340	11694
Japanese	233	12534
Total	677	27507

Publishing data about individuals without revealing sensitive information about them is known to be an important research problem (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006). Now a day, many organizations are increasingly publishing data about their users by anonymizing users' personal information. As a result, the anonymization of data and de-identification techniques are often at the forefront of research institutions' concern when it comes to publishing data. Numerous techniques have been proposed such as l-diversity, k-anonymity, hash function, tokenization etc. for data anonymization and/or pseudonymization of data before publishing. In order to prepare SCROLL dataset, our primary concern was about not to reveal any sensitive information. Therefore, we defined sensitive and non-sensitive information. Sensitive information is those that can uniquely identify a user using the seemingly innocuous attributes gender, date of birth, and postal code etc. Hence, to avoid the identification of these records in dataset, uniquely identifiable information like name, age, email, phone etc. are removed from the table. Moreover, images that contain identify human faces and social situation including a room's interior, laboratory setting, numbers are removed from the table. We did not anonymize locational information. Because we hypothesized that from the coordinate information, top N locations for each user can be measured with different levels of granularity, ranging from a cell sector to whole cell, zip code, city, county, and state (Zang & Bolot, 2011). Table 2 shows the fields of the SCROLL dataset. Note that- (1) any location with NULL refers that GPS signal was not detected hence, we couldn't ensure the study location; (2) empty (i.e. blank) field refers to the information that is intentionally hidden for data privacy.

Table 2: Details of the Fields

Field	Description	Example
Log.Id	Id(anonymized) of each vocabulary log	033d9c9ac6d44b56ac48a47c4600e5e7
Learner.Id	Learner(anonymized) id	ff80818125899fb501258acaadc10001
Default.Lang	Default language of a learner	Chinese, Japanese, and English
Latitude	Latitude of the study location	34.07027
Longitude	Longitude of the study location	134.554844
Create.Time	Creation time of a vocabulary log	6/13/2013 20:09
Image	Image id	114163a52e6545bf8dd8380ceebf9ebd
Vocabulary	Vocabulary that is learned (Bengali)	হাতি (hati)
Meaning	Translation of the vocabulary	象(zo)
Relog	Whether one's log is reclogged by other	033d9c9ac6d44b56ac48a47c4600e5e7
Quiz.date	Quiz was created	6/13/2013 21:19
Quiz.Lat	Latitude where quiz is generated	34.07027
Quiz.Lng	Longitude where quiz is generated	134.554844
Quiz.Content	Content used in a generating quiz	Apple, orange etc.
MyAnswer	Number of options chosen in quiz	1, 2, 3 or 4
Answer	Right answer of the quiz	1, 2, 3 or 4

The dataset is open to the research communities continuously contributing to improve learning analytics, educational data mining, ubiquitous learning, language learning or closely related

research. In order to use the dataset, a researcher is required apply via the project page[1]. Further instructions regarding the dataset including how to download it, data sharing policies, citation information are provided in project's pages[2].

4. Recent Research using the Dataset

The SCROLL dataset is recently used to analyze various problems in educational technology and provide solutions to them. Before 2015, the dataset has been used for personalized knowledge awareness map, computer-supported vocabulary learning, enriching ubiquitous learning experiences, mobile learning, task-based learning, life-logging, dashboard, information visualization, and other related researches. In this section, we articulate some of the recent researches that are carried out with this dataset from 2015 and onwards.

4.1 False-friends Detection across Contexts (Abou-Khalil, Brendan, Flanagan, & Ogata, 2018)

False-friends referred to those words that look alike and pronounced similarly in two languages but differ significantly in meaning. False-friends often create confusion in foreign language learners' minds particularly in informal learning that may lead to miscommunication. Hence, this problem needs to be overcome. In our investigation, utilizing ULLs, the identification of a word's intended meaning under a specific learning context is determined. Then we provided the learner with the appropriate translation, early warnings and quizzes in order to improve the learning process and to avoid false-friends' that s/he might encounter in future.

4.2 Image Recommendation for Informal Vocabulary Learning (Hasnine M. N., Mouri K., Flanagan B., Akcapinar K., Uosaki N., & Ogata H, 2018)

Hasnine et al. proposed Feature-based Context-specific Appropriate Image (FCAI) recommendation system to assist language learners in informal learning of foreign vocabulary. The system is based on the analyzing 25350 ubiquitous learning logs (i.e. a learner's ethnographic information, learning location, learning time, learning context, image information etc.) from the dataset.

4.3 Context-awareness and Personalization (Mouri, Ogata, Uosaki, & Lkhagvasuren, 2016)

Utilizing the ubiquitous learning logs from the dataset, a system is developed that could recommend useful learning logs at the right place in the right time in accordance with personalization of learners. For personalized recommendation, a model named Learner-Kowledge-Place-Time-Experience (LKPTE) is proposed in which 5 parameters, namely word, information, native language, place, and time are used. The finding of this study indicates that the system is capable of increasing learners' learning opportunities.

4.4 Japanese Onomatopoeia Learning (Uosaki, Ogata, Mouri, & Lkhagvasuren, 2015)

In the Japanese language, the number of onomatopoeic words is high. As a result, foreign students learning Japanese language struggle to acquire them. Research, in order to facilitate Japanese onomatopoeia learning, has been carried out using the dataset. A system with the objectives to share onomatopoeia learning contents among onomatopoeia learners, and to organize and reinforce their knowledge is developed. In the evaluation, it was found that the system has high usability and learners' satisfaction because learners got to learn in technology-enhanced learning.

[1] http://eds.let.media.kyoto-u.ac.jp/?page_id=698&lang=en
[2] https://sites.google.com/view/letsscroll/

4.5 Network Visualization (Mouri & Ogata, 2015)

Research on developing innovative visualization system by integrating network visualization technologies with Time-Map based on Ubiquitous Learning Analytics (ULA) is carried out. The SCROLL ULL dataset is used for this research. One of the key focuses of this research was to visualize the relationships between learners and ULLs. With this study, foreign students living in Japan can add their knowledge to the SCROLL system, and then the system can present the learning contents to help them in recalling previously acquired knowledge based on their learning contexts. The paper (Mouri et al., 2015) details the architecture of the network visualization model.

5. Research Scopes

Learning Analytics (LA) and Educational Data Mining (EDM) are emerging interdisciplinary fields that concern with developing innovative methods for exploring unique types of data from educational environments. In recent years, LA and EDM techniques such as statistics, visualization, SNA, concept analysis, classification, clustering, relationship mining and discovery with models are tested in Intelligent Tutoring System(ITS), Technology-enhanced learning, educational technology researches. The scientific research topics shown in Table 3 can be explored (but not limited to) using the SCROLL ubiquitous learning log dataset. Learning analytics, educational data mining, text mining, artificial intelligence, and pedagogical approaches are advised to apply on this dataset.

Table 3: Suggested Topics

Research Scope	Required Baseline Information
Location-based word recommendation	Vocabulary, Place(Lat. and Long.)
Topic generation	Vocabulary, Corpus (Wikipedia, Twitter etc.)
Informal learning behavior analysis	Vocabulary, Time, Place (Lat. and Long.), User profile
Cross-cultural association analysis	Vocabulary, Place (Lat. and Long.), User profile
Data visualization	Vocabulary, Place (Lat. and Long.), User profile
Learning experience transformation	Vocabulary, Log, Meaning (different lang.)
Authentic learning	Vocabulary, Log, Meaning (different lang.)
Data-driven ethnographic research	User profile
Quiz generation	Vocabulary, Log, Quiz info., Quiz history
Context and culture in learning	Vocabulary, User profile, Place (Lat. and Long.)
Adults language learning	Vocabulary, Place (Lat. and Long.), Time
Association analysis	Vocabulary
Image recommendation	Vocabulary, Image info.

For evaluation, log analysis, machine learning, user ratings, self-report, statistical methods, perplexity measurement, network visualization, bleu score, accuracy, recall, precision, correlation measurement, feedback, comparative approach, motivation analysis, effectiveness tests, memory retention etc. are suggested to adapt.

6. Conclusion

Learning analytics and educational data mining are maturing, and have shown great promise for 20th century's education. Some studies have been carried out to accumulate lifelog data of a person's life utilizing scanned material (e.g., articles, books) and digital data recording media (e.g., emails, web pages, phone calls, and digital photos). Analysis of these sort of dataset has gained interest in multimodal learning analytics domain. However, in the context of ubiquitous learning, the representative lifelog dataset where modern machine learning methods can be applied is limited. Indeed, some representative datasets such as USC-HAD dataset, MIT PlaceLab dataset, UC Berkeley's WARD dataset, Microsoft's MyLifeBits dataset, Yamada's Clothes Life Log dataset, Yoshihara's Healthcare Lifelong dataset etc. has drawn the great interest of researchers. In this paper, we describe a dataset consisting of 27507 ubiquitous learning logs and 62466 quiz logs that

are collected from foreign language learners that registered themselves as either Chinese, English or Japanese as their default language. Using SCROLL, we aimed to capture foreign language learning experiences in daily life and reuse them for learning and education. Since SCROLL is intended to be used in general domains and for life-long learning, we will apply it to many application domains including foreign language, math, physics, and science education, and conduct a long-term evaluation with a larger sample of subjects. Another area of our future work is learning analytics. We plan to analyze the accumulated data in the learning logs to find learners' learning patterns and learning habits in order to supply more appropriate learning materials in more appropriate places and at more appropriate times to improve the learning effects of the system. The dataset is prepared by not revealing any sensitive information of the users. Location information is also provided by which top N locations for each user can be measured with different levels of granularity. We also suggest that the data can be segmented time-wise (Year-Month-Week-Day-Hour-Minute-Second), location-wise (i.e. library, café, university, home, city etc.), user profile-wise (nationality, ethnographic background, target language etc.), learning trail-wise etc. For evaluation, log analysis, machine learning, user ratings, self-report, statistical methods, memory retention etc. are recommended. This dataset will be open for the researchers that will take part in the workshop.

Acknowledgements

This research was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (S) Grant Number 16H06304.

References

Abou-Khalil, V., Brendan, Flanagan, & Ogata, H. (2018). Learning false friends across contexts, *Companion Proceedings 8th International Conference on Learning Analytics & Knowledge*, (pp.1-11).

Gurrin, C., Smeaton, A. F., & Doherty, A. R. (2014). LifeLogging: Personal Big Data. *Foundations and Trends in Information Retrieval*, 8(1), 1–125.

Hasnine M. N., Mouri K., Flanagan B., Akcapinar K., Uosaki N., & Ogata H. (2018). Image recommendation for informal vocabulary learning in a context-aware learning environment. *International Conference on Computer in Education(ICCE2018)*. In Press

Hiroaki Ogata, Bin Hou, Mengmeng Li, Noriko Uosaki, Kosuke Mouri, & Songran Liu. (2014). Ubiquitous learning project using life-logging technology in Japan. *Journal of Educational Technology & Society*, 17(2), 85–100.

K. Sakamura, & N. Koshizuka. (2005). Ubiquitous computing technologies for ubiquitous learning. In *IEEE International Workshop on Wireless and Mobile Technologies in Education (WMTE'05)* (pp. 11–20).

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). L-diversity: privacy beyond k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)* (pp.24–36).

Mouri, K., & Ogata, H. (2015). Ubiquitous learning analytics in the real-world language learning. *Smart Learning Environments*, 2(1), 15-33.

Mouri, K., Ogata, H., Uosaki, N., & Lkhagvasuren, E. (2016). Context-aware and personalization method based on ubiquitous learning analytics. *Journal of Universal Computer Science*, 22(10), 1380–1397.

Ogata, H., Li, M., Hou, B., Uosaki, N., El-Bishouty, M. M., & Yano, Y. (2011). SCROLL: Supporting to share and reuse ubiquitous learning log in the context of language learning. *Research & Practice in Technology Enhanced Learning*, 6(2), 69–82.

Uosaki, N., Ogata, H., Mouri, K., & Lkhagvasuren, E. (2015). Japanese onomatopoeia learning support for international students using SCROLL. In *Doctoral Student Consortium (DSC) - Proceedings of the 23rd International Conference on Computers in Education*(ICCE2015), (pp.329–338).

Yahya, S., Ahmad, E. A., & Jalil, K. A. (2010). The definition and characteristics of ubiquitous learning: A discussion. *International Journal of Education and Development Using Information and Communication Technology*, 6(1), 1–12.

Zang, H., & Bolot, J. (2011). Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, (pp. 145–156).

Zhang, M., & Sawchuk, A. A. (2012). USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing(Ubicomp)*, (pp.1036–1043).