MOOC Analytics: Learner Modeling and Content Generation

Guanliang Chen
MOOC Analytics: Learner Modeling and Content Generation

Dissertation

for the purpose of obtaining the degree of doctor at Delft University of Technology
by the authority of the Rector Magnificus prof.dr.ir. T.H.J.J. van der Hagen
Chair of the Board for Doctorates,
to be defended publicly on Monday, 6 May, 2019 at 10:00 AM

by Guanliang CHEN

Master of Software Engineering
South China University of Technology, China
born in Zhanjiang, Guangdong, China.
This dissertation has been approved by the promotors:
Prof.dr.ir. G.J.P.M. Houben
Copromotor: Dr. C. Hauff

Composition of the doctoral committee:
Rector Magnificus chairperson
Prof.dr.ir. G.J.P.M. Houben Delft University of Technology, promotor
Dr. C. Hauff Delft University of Technology, copromotor

Independent members:
Prof.dr. M.M. Specht Delft University of Technology
Prof.dr. D. Gasevic Monash University
Prof.dr. K. Verbert Katholieke Universiteit Leuven
Prof.dr. M. Kalz Heidelberg University of Education
Prof.dr. A. Hanjalic Delft University of Technology

SIKS Dissertation Series No. 2019-13

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems. This work is also supported by the Extension School of Delft University of Technology.

Published and distributed by: Guanliang Chen
E-mail: angus.glchen@gmail.com


Keywords: MOOCs, Learner Modeling, Content Generation, Learning Analytics, Social Web

Copyright © 2019 by Guanliang Chen
All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission of the author.

Cover design by: Longkai Fang and Guanliang Chen.

Printed and bound in The Netherlands by Ipskamp Printing.
Acknowledgments

Upon the completion of my journey of pursuing a PhD degree, I would like to deliver my gratitude to all who helped me overcome the challenges that I encountered in this journey.

First and foremost, I would like to express my highest gratitude to my promotor Geert-Jan Houben, who is smart, kind, and considerate and always cares for his students. Thank you, Geert-Jan! Thanks for having the nice conversation with me when we first met in UMAP 2014 in Denmark. Thanks for providing me with the thoughtful supervision and strong support during the whole period of my PhD study. In particular, thanks for giving me so much useful advice in writing up this thesis. Without your support, I would not have been able to finish this thesis and gain so much on both professional and personal development.

I am deeply indebted to my daily supervisor Claudia Hauff. Thank you, Claudia. It was a great honor to be one of your PhD students and receive the extensive guidance and helpful advice from you, which have made my PhD journey rewarding and enjoyable. It was such a great pleasure working with you. All of the useful research lessons that you taught me are lifetime treasure to me!

I would like to express my gratitude to Marcus Specht, Dragan Gasevic, Katrien Verbert, Marco Kalz, and Alan Hanjalic, for serving as my committee members and providing me with insightful feedback.

I would like to thank my supervisors in my master study, Jian Chen and Li Chen, without whose guidance and encouragement, I would not have the courage to start my PhD journey.

I owe many thanks to my collaborators: Dan Davis, Jun Lin, Tim van der Zee, Markus Krause, Efthimia Aivaloglou, Elle Wang, Luc Paquette, Ioana Jivet, René F. Kizilcec, Yingying Bao, Yue Zhao, Christoph Lofi, Sepideh Mesbah, Manuel Valle Torre, Alessandro Bozzone, Jie Yang, Wenjie Pei, Wing Nguyen, Haoran Xie, and Christopher Brooks. It was truly honorable and enjoyable to work with all of you!
I am grateful to the members of the Web Information Systems (WIS) group and the former WISers for their help and friendship: Marcus Specht, Alessandro Bozzone, Asterios Katsifodimos, Christoph Lofi, Nava Tintarev, Dimitrios Bountouridis, Andrea Mauri, Panagiotis Mavridis, Achilleas Psyllidis, Arthur Camara, Dan Davis, Vincent Gong, Christos Koutras, Sepideh Mesbah, Felipe Moraes, Shabnam Najafian, Jasper Oosterman, Gustavo Penha, Ioannis Petros Samiotis, Shahin Sharifi, Sihang Qiu, Yue Zhao, Carlo van der Valk, Jan Hidders, Stefano Bocconi, Pavel Kucherbaev, Tarmo Robal, Mohammad Khalil, Mónica Marrero, Tamara Brusík, Roniet Sharabi, and Naomi Meyer. I cannot thank Sepideh, Shahin, Sihang, Yue, Shabnam, and Felipe enough for the enjoyable time we had together. I would like to give special thanks to Jie, Yue, Sihang, and Sepideh. Thanks, Jie, for the kind introduction about our research group as well as TU Delft when we first met and helping me prepare for the PhD application. Thanks, Yue, thanks for helping me adapt to life in the Netherlands and doing me countless favors in the past four years. Thanks, Sihang, for helping me manage stuff related to my thesis and defense. Thank you so much, Sepideh! I really appreciate your help and support for my PhD study and I will always cherish the days when we were officemates.

I would like to thank the friends that I made during my PhD study: Shan-shan Ren, Zhe Hou, Wei Dai, Shengzhi Xu, Jian Fan, Yingying Bao, Zhengwu Huang, Yan Song, Jun Lin, Jingyang Liu, Chen Huang, Kai Wu, Jiaye Liu, and Jiahao Lu. In particular, I would like to thank Jiahao. Thanks, Jiahao, for the delicious meals that you prepared for me, the help you provided with me during the last four months of my PhD study, and the enjoyable time that we spent together in exploring Europe. It was a great pleasure to be your friend.

I owe many thanks to my friends in China: Linya Zhang, Haijing Zhang, Haojun Chen, Yin Ye, and Xiaoli Huang, for bringing enormous joy to my life. Without your company, my life would not be so colorful! I would like to give special thanks to Linya, who always offers me generous help and support whenever I need. Thanks, Linya!

Last but not least, I would like to express my deepest gratitude to my parents, Jiadong Chen and Qunzhen Su, for their never-ending love, encouragement, and support. Also, I would like to thank my dear sister and brother-in-law, Hong Chen and Yong Wang, for their unconditional caring and guidance. Particularly, I would like to express my gratitude to my adorable nephews, Weiming Wang and Yining Wang, whose smile helped me overcome many difficulties that I had in this four-year journey. (最后，我想向我的父母陈家东与陈群珍致以我衷心的感谢，谢谢他们对我永不停歇的爱，鼓励与支持。)
Acknowledgments

同样，我想向我亲爱的姐姐陈虹与姐夫王勇表达感谢，谢谢他们对我无条件的关怀与教导。特别地，我想谢谢我两个可爱的小外甥王伟茗与王奕宁，你们可爱的笑脸帮助我克服了过去四年遇到的许多困难。)

Guanliang Chen
March 2019
Melbourne, Australia
# Contents

1 Introduction ............................................. 1
   1.1 Motivation and Objectives .......................... 1
   1.2 Research Questions and Contributions .............. 3
   1.3 Thesis Outline and Origin of Chapters .............. 8

2 Learner Identification across Social Web Platforms ...... 9
   2.1 Introduction ......................................... 10
   2.2 Social Web & MOOCs .................................. 12
   2.3 Approach ............................................. 13
      2.3.1 Locating Learners on the Social Web ........... 13
      2.3.2 Social Web Platforms .............................. 14
      2.3.3 Social Web Data Analysis ........................ 15
   2.4 MOOC Learners & the Social Web ..................... 19
   2.5 Results ............................................... 21
      2.5.1 Learners on Twitter ............................... 21
      2.5.2 Learners on LinkedIn ............................... 23
      2.5.3 Learners on StackExchange ......................... 28
      2.5.4 Learners on GitHub ................................. 30
   2.6 Conclusion ........................................... 31

3 Learning Transfer ........................................ 33
   3.1 Introduction ......................................... 34
   3.2 Background ......................................... 36
   3.3 FP101x ............................................. 38
   3.4 Methodology ......................................... 38
      3.4.1 Research Hypotheses .............................. 38
## Contents

### 6 LearningQ for Educational Question Generation

6.1 Introduction ........................................ 104
6.2 Related Work ........................................ 107
   6.2.1 Question Generation ........................... 107
   6.2.2 Datasets for Question Generation ............. 108
   6.2.3 Question-worthy Sentence Selection .......... 109
6.3 Data Collection ....................................... 110
   6.3.1 Data Sources .................................. 110
   6.3.2 Question Classification for Khan Academy .......... 111
   6.3.3 Final Statistics of LearningQ .................. 113
6.4 Sentence Selection Strategies ......................... 115
6.5 Data Analysis on LearningQ .......................... 116
   6.5.1 Document & Question Lengths ............... 117
   6.5.2 Topics, Interrogative Words, and Readability ...... 117
   6.5.3 Cognitive Skill Levels ...................... 120
6.6 Experiments and Results .................................. 123
   6.6.1 Experimental Setup .......................... 124
   6.6.2 Evaluation on LearningQ .................. 127
   6.6.3 Evaluation on Sentence Selection Strategies .... 129
6.7 Conclusion ......................................... 132

### 7 Conclusion

7.1 Summary of Contributions .......................... 134
7.2 Future Work ........................................ 137
   7.2.1 Adaptive Learning in MOOCs .................. 137
   7.2.2 Interactive Learning in MOOCs .......... 138
   7.2.3 Content Enrichment in MOOCs .................. 139

### Bibliography

141

### List of Figures

159

### List of Tables

163

### Summary

167

### Samenvatting

171
Contents

Curriculum Vitae 175
Chapter 1

Introduction

1.1 Motivation and Objectives

*Lifelong learning* has been widely recognized as an important social issue. As indicated by UNESCO, one objective stated in the Education 2030 Framework for Action is to “promote lifelong learning as the leading educational paradigm for achieving inclusive and sustainable learning societies” [153]. Learning used to be largely restricted to formal education in schools. With the development of technology, people now have more options to receive education and learn. Massive Open Online Courses (MOOCs), as one of the available options, are endowed with the mission to *educate the world* [121]. MOOCs refer to online courses that are designed for an unlimited number of participants. In MOOCs, the learning materials are distributed over the Web, which can be accessed by learners with internet connections anytime and anywhere [112]. There are two types of MOOC platforms: *topic-agnostic* and *topic-specific*. Topic-agnostic platforms (e.g., edX and Coursera) provide courses covering a wide range of topics, while *topic-specific* MOOC platforms (e.g., Duolingo and Codeacademy) focus on courses in one specific topic. UNESCO regards MOOCs as an essential tool to “promote lifelong learning opportunities for all” [121]. In fact, MOOCs are becoming increasingly popular. According to Class Central [141], by the end of 2017, there have been more than 81 million learners enrolled in 9,400 MOOCs in 33 MOOC platforms including edX, Coursera, etc.

1https://www.edx.org/
2https://www.coursera.org/
3https://www.duolingo.com/
4https://www.codecademy.com/
To better support MOOC learners, there have been many works on investigating MOOC learning. Typically, these works employed the data traces generated by learners within MOOC platforms to investigate their behavior during the running of a course, such as course navigation patterns of learners of various demographics [63], the impact of different video types on learner engagement [66], the sentiment expressed by learners in forum posts [163], the effect of instructor involvement [149]. Still, there are many other aspects of MOOC learning to be explored.

In this thesis, we focus on (i) learner modeling and (ii) generation of educational material for both topic-agnostic and topic-specific MOOC platforms. For learner modeling in the topic-agnostic platforms, as there have been a lot of works utilizing the learner traces generated within the MOOC platforms, we hypothesize that we can better understand learners by moving beyond the MOOC platforms and exploring other data sources on the wider Web, especially the Social Web. Nowadays, hundreds of millions of users are heavily using Social Web platforms with different purposes, such as microblogging (Twitter), professional networking (LinkedIn), Q&A (StackExchange) and collaborative programming (GitHub). Previous research demonstrated that abundant data traces in the Social Web platforms can be used to reveal detailed information about users such as age [113], occupation [127], language proficiency [159], and professional experience [26]. Therefore, we investigate what attributes can be revealed for modeling MOOC learners with the aid of the Social Web, not only during a MOOC but also before and after the MOOC. With regard to the topic-specific MOOC platforms, given that there are only a few works on modeling learners [63, 66, 163], we investigate what approaches can be used to enable a better understanding of learners in these platforms.

For generation of educational material, previous research demonstrated that certain Social Web data (e.g., code snippets in GitHub, Q&A pairs in StackExchange) can be reused by users of similar interests and needs [111, 110]. Therefore, we investigate what Social Web data can be used to generate educational material and potentially benefit MOOC learners.
1.2 Research Questions and Contributions

In the following, we present the research questions investigated in Chapters 2-6. Each chapter explores different data sources (MOOC platforms and Social Web platforms), and focuses on different MOOC stages (before, during and after a MOOC), which are summarized in Figure 1.1. In total, we consider two MOOC platforms and eight Social Web platforms in our thesis, and most of our works focus on the stage of during a MOOC.

<table>
<thead>
<tr>
<th>MOOC Stages</th>
<th>Data Sources</th>
<th>Before</th>
<th>During</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOOC Platforms</td>
<td>edX</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>duolingo</td>
<td></td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Social Web Platforms</td>
<td>Gravatar</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>twitter</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LinkedIn</td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>StackExchange</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>GitHub</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Upwork</td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KHAN</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TEDx</td>
<td></td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.1: An overview of the MOOC stages and data sources investigated in Chapters 2-6. The number in a cell represents the corresponding chapter, which focuses on the MOOC stage specified in the column and the MOOC platform or the Social Web platform specified in the row.

To explore whether the Social Web can be used to enable learner modeling beyond the topic-agnostic MOOC platforms, in Chapter 2 we first verify whether MOOC learners are active on Social Web platforms and investigate how to reliably identify these learners. As people tend to be attracted by different Social Web platforms and correspondingly leave various data traces in those platforms, it is a non-trivial task to identify MOOC learners across multiple platforms and further gather information relevant to their learning activities in MOOCs. Specifically, we investigate the following research questions:
RQ 1.1 On what Social Web platforms can a significant fraction of MOOC learners be identified?

RQ 1.2 Are learners who demonstrate specific sets of traits on the Social Web drawn to certain types of MOOCs?

RQ 1.3 To what extent do Social Web platforms enable us to observe (specific) user attributes that are highly relevant to the online learning experience?

To answer those questions, we consider over 320,000 learners from eighteen MOOCs in edX and propose a systematic methodology to reliably identify these learners across five popular Social Web platforms, i.e., Gravatar, Twitter, LinkedIn, StackExchange and GitHub. Furthermore, we explore what valuable data traces can be gathered from the considered platforms and used to investigate MOOC learning. In particular, we find that over one-third of learners from a MOOC teaching functional programming are actively engaged with GitHub, the most popular social coding platform in the world to date, and have left abundant coding traces in the specific platform. More importantly, this enables a first investigation on learning transfer, which refers to the application of knowledge or skills gained in the learning environment to another context.

Based on the observation of the active engagement of learners from a programming MOOC in GitHub in Chapter 2, in Chapter 3, we zoom in on the coding traces of these learners in GitHub and continue the investigation of their learning transfer, as a perspective to examine the influence of the course on the learners. Concretely, we investigate the following research questions:

RQ 2.1 To what extent do learners from a programming MOOC transfer the newly gained knowledge to practice?

RQ 2.2 What type of learners are most likely to make the transfer?

RQ 2.3 How does the transfer manifest itself over time?

To answer those questions, we conduct a longitudinal analysis on both the MOOC platform data and the GitHub data. We find that only a small fraction of engaged learners (about 8%) display transfer. To our knowledge, this analysis has been the first to introduce the use of the Social Web to model learners’ knowledge application beyond the learning platform.
1.2. Research Questions and Contributions

As indicated before, only a few works attempted to model learners in topic-specific MOOC platforms like Duolingo and Codeacademy. The main reason for this is the lack of public available datasets from these platforms to enable further research. In the Second Language Acquisition Modeling challenge \cite{140} organized by Duolingo, which is the largest language-learning MOOC platform in the world, three large-scale datasets collected from its learners over the first 30 days of language learning were released. With the datasets, we are able to gain more insights about learners in topic-specific MOOC platforms.

In Chapter \ref{chap:3}, we use the three released datasets to analyze learners’ behavior in Duolingo and model their mastery of the taught knowledge over time (i.e., knowledge tracing \cite{125}). Concretely, we investigate the following research question:

**RQ 3.1** What factors are correlated with learners’ language learning performance?

To answer the question, we analyze the three Duolingo datasets to identify a range of features that are correlated with learners’ performance and further investigate their effectiveness in predicting learners’ future performance. We demonstrate that the learning performance, which is measured by learners’ accuracy in solving exercises and the amount of vocabulary learned, is correlated with not only learners’ engagement with a course but also contextual factors like the devices being used.

In Chapter \ref{chap:3}, we have shown that learners transfer the acquired knowledge to practice. In Chapter \ref{chap:4}, we investigate whether learners could apply the acquired knowledge to solve real-world tasks, i.e., paid tasks which are retrieved from online marketplaces and can be solved by applying the knowledge taught in a course. If learners are able to solve such tasks, ultimately, we envision a recommender system that presents learners with paid relevant tasks from online marketplaces. By solving these tasks, learners, who cannot spend a large amount of time in learning because of the need to work and earn a living, could earn money and thus gain more time for learning with the MOOC. To investigate the feasibility of the proposed recommender system, we investigate the following questions:

**RQ 4.1** Are MOOC learners able to solve real-world (paid) tasks from an online work platform with sufficient accuracy and quality?

**RQ 4.2** How applicable is the knowledge gained from MOOCs to paid tasks offered by online work platforms?
RQ 4.3 To what extent can an online work platform support MOOC learners (i.e., are there enough tasks available for everyone)?

RQ 4.4 What role do real-world (paid) tasks play in the engagement of MOOC learners?

To answer those questions, we consider a MOOC teaching data analysis in edx and manually select a set of paid tasks from Upwork\(^{10}\), one of the most popular freelancing marketplaces in the world, and present the selected tasks to learners and observe how learners interact with these real-world tasks. We find that these tasks can be solved by MOOC learners with high accuracy and quality. This demonstrates the potential of using freelancing paid tasks to enrich MOOC content.

Questions are recognized as essential not only for assessment but also for learning because questions allow learners to not only assess their understanding of concepts but also to reflect on their knowledge state and then better direct their learning efforts\(^ {8, 128}\). However, designing a suitably large question bank to meet the needs of MOOC learners is a time-consuming and cognitively demanding task for course instructors. To ease the burden of the instructors, automatic question generation has been proposed and investigated by researchers to automate the question creation process with the aid of machine learning techniques\(^ {69, 110, 136}\). Ideally, we can construct a question generator, which takes an article of any learning topic as input and generates a set of questions that are relevant to the article and useful for assessment or discussion. To this end, two challenges need to be overcome. Firstly, a large-scale dataset covering questions of various cognitive levels from a set of diverse learning topics should be collected. With the collected dataset, we are able to discover common question-asking patterns and inform the construction of the question generator. Secondly, given that an article often contains a limited number of sentences that are worth asking questions about, i.e., those carrying important concepts, we need to develop effective strategies to identify question-worthy sentences from the article before using them as input to the question generator. To deal with the challenges, we turn to education-oriented Social Web platforms (e.g., TED-Ed\(^ {11}\), Khan Academy\(^ {12}\), Codecademy) because these platforms typically have accumulated a substantial amount of high-quality questions generated by instructors and learners. Therefore, in Chapter Chapter 6, we investigate whether we can

\(^{10}\text{https://www.upwork.com/}
\(^{11}\text{https://ed.ted.com/}
\(^{12}\text{https://www.khanacademy.org/}

use the education-oriented Social Web platforms to collect a large-scale educational question dataset and further use the dataset to develop effective strategies to identify question-worth sentences from an article. Correspondingly, we investigate the following research questions:

RQ 5.1 Can a large-scale and high-quality educational question dataset be collected from the Social Web?

RQ 5.2 What are effective strategies in identifying question-worthy sentences from an article?

To answer those questions, we rely on TED-Ed and Khan Academy to retrieve an educational question dataset, LearningQ, which contains over 230K document-question pairs generated by both instructors and learners. To the best of our knowledge, LearningQ is the largest dataset that can be used for educational question generation. We demonstrate that LearningQ consists of high-quality questions covering not only all cognitive levels in the Bloom’s Revised Taxonomy [104] but also various learning topics. We show that it is a challenging task to automatically generate educational questions, even with sufficient training data and state-of-the-art question generation techniques. Besides, we develop and compare a total of nine strategies to select question-worthy sentences from an article and demonstrate that questions in learning contexts usually are based on source sentences that are informative, important, or contain novel information.

In summary, this thesis makes the following research contributions.

• We contribute a systematic methodology to reliably identify learners across five popular Social Web platforms and derive a set of valuable learner attributes to investigate MOOC learning.

• We contribute a novel approach to use GitHub to complement data traces within MOOC platforms as a means to investigate learner behavior (i.e., learning transfer) beyond the MOOC platform.

• We contribute an analysis to identify factors (e.g., learners’ engagement with a course, the learning devices being used) that are related to learners’ performance in second language acquisition.

• We contribute a study to demonstrate that learners can apply the knowledge acquired from a MOOC to solve real-world tasks with high accuracy and quality.
• We contribute a large educational dataset (LearningQ) for automatic question generation and investigate nine strategies in selecting question-worthy sentences from an article.

1.3 Thesis Outline and Origin of Chapters

This thesis consists of seven chapters. The current chapter describes the motivation, objectives, and research questions as well as contributions. All the main chapters (Chapter 2–6) are based on full research papers published in conferences or journals, except for Chapter 4, which is published as a workshop paper.

• **Chapter 2** is based on the paper published at the ACM Conference on Web Science [31].

• **Chapter 3** is based on the paper published at the ACM Conference on Learning at Scale [30], where the paper received the Honorable Mention Award.

• **Chapter 4** is based on the paper published at the Workshop on Innovative Use of NLP for Building Educational Applications [32].

• **Chapter 5** is based on the paper published at the IEEE Transactions on Learning Technologies [28].

• **Chapter 6** is based on the paper published at the International AAAI Conference on Web and Social Media [33] and includes new research work.

Lastly, **Chapter 7** concludes this thesis by summarizing the main findings and contributions. Furthermore, we provide an outlook on future research directions in relevant fields.
Chapter 2

Learner Identification across Social Web Platforms

In this chapter, we first conduct an exploratory study to verify whether MOOC learners are active in the Social Web and how to reliably identify their accounts across various Social Web platforms. This study is intended to serve as a foundation to collect learner traces beyond the MOOC platform and investigate questions that cannot be answered by solely utilizing the data traces learners leave within the MOOC platform. To this end, we consider over 320,000 learners from eighteen MOOCs in edX. Notice that not every Social Web platform attracts a large number of learners and is open for user identification and data retrieval, we eventually consider five popular Social Web platforms in our study, i.e., Gravatar, Twitter, LinkedIn, StackExchange and GitHub. Furthermore, we investigate what data traces can be collected from these platforms and used to derive learner attributes that are relevant to their learning activities in the MOOC setting. The contributions of this chapter have been published in [31].
Chapter 2. Learner Identification across Social Web Platforms

2.1 Introduction

Online education recently entered a new era of large-scale, free and open-access which has revolutionised existing practices. This new era dates from 2011, when the University of Stanford released its initial three MOOCs. Today, a wide range of courses in the humanities, business and natural sciences are offered for free with millions of learners taking advantage of them.

At the same time, however, the initial predictions of the “MOOC revolution” (universities will become obsolete) have not come to pass. On the contrary, MOOCs today generally suffer from a lack of retention — many learners sign up, but on average less than 7% complete a course.

Examining the current nature of MOOCs reveals an important clue as to why they, as yet, fail to realize their full potential. Although the “MOOC revolution” changed online education with respect to scale and openness, it did not involve any truly novel pedagogical approaches or education technologies. Currently, many MOOCs revolve around a set of videos, a set of quizzes and little else (the so-called “xMOOCs”). Instead, new approaches are necessary that support learning under the unique conditions of MOOCs: (i) the extreme diversity among learners (who come from diverse cultural, educational and socio-economic backgrounds), and, (ii) the enormous learner-staff ratio, which often exceeds 20,000:1.

In order to improve the learning experience and retention, MOOC data traces (i.e. learners’ clicks, views, assignment submissions and forum entries) are being employed to investigate various aspects of MOOC learning, such as the effect of lecture video types on learner engagement, the introduction of gamification, the impact of instructor involvement and the significance of peer learning.

Few data-driven research works go beyond the data learners generate within a MOOC platform. We argue that we can potentially learn much more about MOOC learners if we move beyond this limitation and explore the learners’ traces on the wider Web, in particular the Social Web, to gain a deeper understanding of learner behavior in a distributed learning ecosystem. Hundreds of millions of users are active on the larger Social Web platforms such as Twitter and existing research has shown that detailed user profiles can be built from those traces, covering dimensions such as age, interests, personality, location and occupation.

While MOOC learners are usually invited to participate in pre-course surveys that include inquiries about their demographics and motivations, not
2.1. Introduction

all of them do (and those who do may fill in non-credible or false information), with return rates hovering around 10%.\footnote{An estimate we derived based on the MOOCs we consider in this work. This percentage drops to 1% or less when considering post-course surveys, i.e. questionnaires conducted at the end of a MOOC.} In addition, these surveys can only provide a very limited view of the learners as the return rate drops with every question that is added to the questionnaire and, finally, questionnaires offer us only a snapshot-based perspective as learners cannot be polled continuously across a period of time.

We hypothesize that the Social Web can provide us with a source of diverse, fine-grained and longitudinal learner traces we can exploit in order to (i) derive more extensive learner profiles for a larger learner population than is possible through pre/post-MOOC surveys, and, (ii) investigate questions that cannot be investigated solely based on the traces learners leave within MOOC environments (e.g. the uptake of learned concepts in practice).

In this work we provide a first exploratory analysis of more than 329,000 MOOC learners and the Social Web platforms they are active on, guided by the following three Research Questions:

RQ 1.1 On what Social Web platforms can a significant fraction of MOOC learners be identified?

RQ 1.2 Are learners who demonstrate specific sets of traits on the Social Web drawn to certain types of MOOCs?

RQ 1.3 To what extent do Social Web platforms enable us to observe (specific) user attributes that are highly relevant to the online learning experience?

Our contributions can be summarized as follows:

• We provide a methodology to reliably identify a subset of learners from a set of five Social Web platforms and eighteen MOOCs. Depending on the MOOC/platform combination, between 1% and 42% of the learners could reliably be identified.

• We show that it is indeed possible to derive valuable learner attributes from the Social Web which can be used to investigate learner experience in MOOCs.
• We show that the tracking of learners over time (in the case of GitHub we consider three years of data traces) enables us to investigate the impact of MOOCs in the long-term.

2.2 Social Web & MOOCs

The wider Web is starting to be viewed as a source of useful information in MOOC learning analytics — the field concerned with the understanding and optimization of learning in massive open online courses. Existing works focus on the analysis of Social Web platforms during the running of a MOOC in order to learn more about the interactions and processes occurring within a MOOC. These analyses are not conducted on the individual learner level, but on the aggregated group level, without the explicit matching of MOOC learners to Social Web profiles.

Alario et al. [5] investigate the learners’ engagement with two built-in MOOC platform components (Q&A and forum) and three external Social Web portals (Facebook, Twitter and MentorMob) during the running of a single MOOC. Learners’ MOOC and Social Web identities are not matched directly, instead, learners are asked to join a specific Facebook group and use a course-specific Twitter hashtag. The authors find that despite the active encouragement of the platforms’ usage to exchange ideas and course materials, after the initial phase of excitement, participation quickly dropped off. Similarly, van Treeck & Ebner [156] also rely on Twitter hashtags to identify the microblog activities surrounding two MOOCs. They (qualitatively) analyse the tweet topics, their emotions and the extent of actual interactions among learners on Twitter and find a small group of MOOC participants (6%) to have generated more than half of all microblog content.

Garcia et al. [59] analysed the concurrent Twitter activities of students taking a ‘Social Networking and Learning’ MOOC to track their engagement and discussion beyond the MOOC environment by designating and tracking hashtagged conversation threads. In the same MOOC, [42] presented a generalisable method to extend the MOOC ecosystem to the Social Web (in this case Google+ and Twitter) to both facilitate and track students’ collaborations and discussions outside of the immediate context of the course.

[50] tracked Twitter interactions among MOOC students to understand the dynamics of social capital within a connectivist MOOC environment, which is inherently decentralised and distributed across platforms. This work was primarily concerned with learner-learner relationships in the
context of a MOOC—not individual learner traits. And, more broadly, [SI] explored the types and topics of conversations that happen in the Social Web concurrent to a MOOC.

Four observations can be made based on these studies: existing works (i) analyze one or two Social Web platforms only, (ii) are usually based on experiments within a single MOOC, (iii) do not require a learner identification step (as an intermediary such as a Twitter hashtag is employed), and (iv) focus on learner activities exhibited during the running of a MOOC that are topically related to the MOOC content (e.g. ensured through the use of moderated Facebook group).

In contrast, we present a first exploratory analysis across eighteen MOOCs and five Social Web platforms exploring the learners’ behaviours, activities and created content over a considerably longer period of time.

2.3 Approach

In this section, we first describe our three-step approach to locate a given set of MOOC learners on Social Web platforms, before going into more detail about the analyses performed on the learners’ Social Web traces.

2.3.1 Locating Learners on the Social Web

On the edX platform, a registered learner $l_i$ is identified through a username, his or her full name and email address (as required by the platform), i.e. $l_i = (\text{login}_i, \text{name}_i, \text{email}_i)$. On a Social Web platform $P_j$, the publicly available information about a user $u^j$ usually consists of (a subset of) username, full name, email address and profile description. The profile description is often semi-structured and may also contain links to user accounts on other Social Web Platforms $P_2, \ldots, P_z$. A common assumption in this case (that we employ as well) is that those accounts belong to the same user $u$.

For each Social Web platform $P_j$ we attempt to locate $l_i$ through a three-step procedure:

**Explicit** If $P_j$ enables the discovery of users via their email address, we use email$_i$ to determine $l_i$’s account $u^j_i$ on $P_j$. If available, we also crawl the profile description of $u^j_i$, the profile image (i.e. the user avatar) and extract all user account links to other Social Web platforms under the assumption stated before.
Chapter 2. Learner Identification across Social Web Platforms

**Direct** This step is only applied to the combination of learners and Social Web platforms \((l_i, P_j)\) for which no match was found in the Explicit step. We now iterate over all extracted account links from the Direct step and consider \(l_i\)'s account on \(P_j\) to be found if it is in this list.

**Fuzzy** Finally, for pairs \((l_i, P_j)\) not matched in the Direct step, we employ fuzzy matching: we rely on \(l_i\)'s name \(_i\) & login \(_i\) and search for those terms on \(P_j\). Based on the user (list) returned, we consider a user account a match for \(l_i\), if one of the following three conditions holds:

(i) the profile description of the user contains a hyperlink to a profile that was discovered in the Explicit or Direct step,

(ii) the avatar picture of the user in \(P_j\) is highly similar to one of \(l_i\)'s avatar images discovered in the Explicit or Direct step (we measure the image similarity based on image hashing \([143]\) and use a similarity threshold of 0.9), or,

(iii) the username and the full name of the user on \(P_j\) and \(l_i\) are a perfect match.

### 2.3.2 Social Web Platforms

Our initial investigation focused on ten globally popular Social Web platforms, ranging from Facebook and Twitter to GitHub and WordPress. We eventually settled on five platforms, after having considered the feasibility of data gathering and the coverage of our learners among them. Concretely, we investigate the following platforms:

**Gravatar** is a service for providing unique avatars to users that can be employed across a wide range of sites. During our pilot investigation, we found Gravatar to be employed by quite a number of learners in our dataset. Given that Gravatar allows the discovery of users based on their email address, we employ it as one of our primary sources for Explicit matching. We crawled the data in November 2015. We were able to match 25,702 edX learners on Gravatar.

**StackExchange** is a highly popular community-driven question & answering site covering a wide range of topics. The most popular sub-site on this platform is StackOverflow, a community for computer programming related questions. StackExchange regularly releases a full “data dump” of their content that can be employed for research purposes. We employed the data
release from September 2015 for our experiments. We were able to match 15,135 edX learners on StackExchange.

**LinkedIn** is a business-oriented social network users rely on to find jobs, advertise their skill set and create & maintain professional contacts. The public profiles of its users can be crawled, containing information about their education, professional lives, professional skills and (non-professional) interests. We crawled the data in November 2015. We were able to match 19,405 edX learners on LinkedIn.

**Twitter** is one of the most popular microblogging portals to date, used by hundreds of millions of users across the globe. Twitter allows the crawling of the most recent 3,200 tweets per user. We crawled the data in December 2015 and January 2016. We were able to match 25,620 edX learners on Twitter.

**GitHub** is one of the most popular social coding platforms, allowing users to create, maintain and collaborate on open-source software projects. The GitHub platform creates a large amount of data traces, which are captured and made available for research through two large initiatives: GitHub Archive⁷ and GHTorrent⁸. For our work, we rely on all data traces published between January 1, 2013 and December 31, 2015. We were able to match 31,478 edX learners on GitHub.

In addition, we are interested in how many learners are observed across more one platform. The numbers of learners that can be matched across 2, 3, 4, 5 platforms are 14824, 6980, 3129, 1125, respectively.

### 2.3.3 Social Web Data Analysis

As our work is exploratory in nature, we employ a range of data analysis approaches that enable us to explore our gathered data traces from various angles.

**t-SNE.** Many of our user profiles are high-dimensional: a LinkedIn user may be represented through a vector of his or her skills⁹, and a Twitter user profile may be encoded as a vector of the entities or hyperlinks mentioned in his or her tweets. If we are interested to what extent those user profiles

---

⁷ [https://www.githubarchive.org/](https://www.githubarchive.org/)
⁸ [http://ghtorrent.org/](http://ghtorrent.org/)
⁹ The dimension of the vector space depends on the number of unique skills in the dataset, with a single skill being encoded in binary form.
are similar or dissimilar for users (learners) that are taking different kinds of MOOCs, we can visualize these similarities using t-SNE (t-Distributed Stochastic Neighbor Embedding \[154\]), a visualization approach for high-dimensional data that computes for each datapoint a location on a 2D (or 3D) map. t-SNE\[10\] creates visualizations that reveal the structure of the high-dimensional data at different scales and has been shown to be superior to related non-parametric visualizations such as Isomaps \[3\].

**Age and gender prediction.** Predicting certain user attributes based on a user’s Social Web activities is an active area of research. It has been shown that attributes such as age \[113\], gender \[11\], personality \[79\], home location \[106\] and political sentiments \[14\] (to name just a few) can be predicted with high accuracy from Social Web data sources.

In our work we focus on the prediction of age and gender, as those two attributes can be inferred of Social Web users with high accuracy. We also have intuitions concerning the age and gender (in contrast to, for instance, their personalities) of the learners that take our MOOCs (e.g. a computer science MOOC is likely to have a larger pool of male participants), enabling us to judge the sensibility of the results.

The main challenge in this area of work is the collection of sufficient and high-quality training data (that is, Social Web users with known age, gender, location, etc.). Once sufficient training data has been obtained, standard machine learning approaches are usually employed for training and testing.

In our work, we make age and gender predictions based on tweets and employ the models provided by \[139\] who utilized the English language Facebook messages of more than 72,000 users (who collectively had written more than 300 million words) to create unigram-based age & gender predictors based on Ridge regression \[77\]. The age model $M_{age}$ contains 10,797 terms and their weights $w_t$. To estimate the age of a user $u$, we extract all his English language tweets (excluding retweets), concatenate them to create a document $D_u$ and then employ the following formulation:

$$age_u = w_0 + \sum_{t \in M_{age}} w_t \times \frac{freq(t, D_u)}{|D_u|}. \tag{2.1}$$
2.3. Approach

Here, $|D_u|$ is the number of tokens in $D_u$, $w_0$ is the model intercept and $\text{freq}(t, D_u)$ is the term frequency of $t$ in $D_u$. Only terms in $D_u$ that appear in $M_{\text{age}}$ have a direct effect on the age estimate. The model is intuitively understandable; the five terms with the largest positive weights (indicative of high age) are {grandson, daughter, daughters, son, folks}. Conversely, the five terms with the largest negative weights (indicative of a young user) are {parents, exams, pregnant, youth, mommy}.

The gender prediction is derived in an analogous fashion based on model $M_{\text{gender}}$, which consists of 7,137 terms and their weights. In contrast to the age estimation (which provides us with a continuous estimate), we are interested in a binary outcome. Thus, after the regression stage, classification is performed: if the estimation is $\geq 0$, the user is classified as female and otherwise as male. Once more, the model is intuitive; the largest negative weights (indicating maleness) are {boxers, shaved, haircut, shave, girlfriend}.

**Learning Transfer.** Existing investigations into student learning *within* MOOC environments are commonly based on pre- & post-course surveys and log traces generated within those environments by the individual learners [74]. With a crude, binary measure of learning, the success (pass/no-pass) of the learner could be labeled. While learning is an important success measure, we also believe that the amount of *learning transfer* [94] that is taking place should be considered: do learners actually utilize the newly gained knowledge in practice? Are learners expanding their knowledge in the area over time or do they eventually move back to their pre-course knowledge levels and behaviours? While most Social Web platforms do not offer us insights into this question, for MOOCs (partially) concerned with the teaching of programming languages (such as *Functional Programming*) we can rely on the *GitHub* platform to perform an initial exploration of this question.

*GitHub* provides extensive access to data traces associated with *public* coding repositories, i.e. repositories visible to everyone. [12]. *GitHub* is built around the *git* distributed revision control system, which enables efficient distributed and collaborative code development. *GitHub* not only provides relevant repository metadata (including information on how popular a repository is, how many developers collaborate, etc.), but also the actual code that was altered. As the *GitHub Archive* makes all historic *GitHub* data traces easily accessible, we relied on it for data collection and extracted all *GitHub*

---

12 Data traces about private repositories are only available to the respective repository owner.
13 [https://www.githubarchive.org/](https://www.githubarchive.org/)
data traces available between January 1, 2013 and June 30, 2015 (five months after the end of the programming MOOC in our dataset). We then filtered out all traces that were not created by the 31,478 learners we identified on the GitHub platform. Of the more than 20 GitHub event types\textsuperscript{14}, we only consider the PushEvent as vital for our analysis.

Every time code is being updated (“pushed” to a repository), a PushEvent is triggered. Figure 2.1 contains an excerpt of the data contained in each PushEvent. The most important attributes of the event are the created_at timestamp (which allows us to classify events as before/during/after the running of the programming MOOC), the actor (the user doing the “push”) and the url, which contains the URL to the actual diff file. While the git protocol also allows a user to “push” changes by another user to a repository (which is not evident from inspecting the diff file alone), this is a rare occurrence among our learners: manually inspecting a random sample of 200 PushEvents showed 10 such cases.

A diff file shows the difference between the last version of the repository and the new one (after the push) in terms of added and deleted code. For each of the identified PushEvents by our learners, we crawled the corresponding diff file, as they allow us to conduct a more fine-grained code analysis. As a first step in this direction, we identified the number of additions and deletions a user conducts in each programming language based on the filename extensions found in the corresponding diff file.

\textsuperscript{14}https://developer.github.com/v3/activity/events/types/
2.4 MOOC Learners & the Social Web

As a starting point for our investigation we utilize eighteen MOOCs that have run between 2013 and 2015 on the edX platform — the largest MOOCs conducted by the Delft University of Technology (situated in the Netherlands) to date; the courses cover a range of subjects in the natural sciences, computer science and the humanities and were all taught in English. An overview of the MOOCs can be found in Table 2.1: we deemed the MOOC titles not to be self-explanatory, so we also added the MOOC’s “tag line”. Apart from the *Pre-university Calculus* (specifically geared towards pre-university learners) and the *Topology in Condensed Matter* (aimed at MSc and PhD physics students) courses, the MOOCs were created with a wide variety of learners in mind. All courses follow the familiar MOOC recipe of weekly lecture videos in combination with quizzes and automatically (or peer-) graded assignments.

The MOOCs vary significantly in size. The largest MOOC (*Solar Energy 2013*) attracted nearly 70,000 learners, while the smallest one (*Topology in Condensed Matter 2015*) was conducted with approximately 4,200 learners. While the majority of learners register for a single MOOC only, a sizable minority of learners engage with several MOOCs and thus the overall number of unique learners included in our analysis is 329,200.

To answer RQ 1.1, Table 2.1 summarizes to what extent we were able to identify learners across the five Social Web platforms, employing the three-step procedure described in Section 2.3.1. Note that the numbers reported treat each course independently, i.e. if a learner has registered to several courses, it will count towards the numbers of each course.

The percentage of learners we identify per platform varies widely across the courses between 4-24% (*Gravatar*), 1-22% (*StackExchange*), 3-42% (*GitHub*), 4-11% (*LinkedIn*) and 5-18% (*Twitter*) respectively. *Functional Programming* is the only MOOC we are able to identify more than 10% of the registered learners across all five Social Web platforms. While this finding by itself is not particularly surprising — two of the five Social Web platforms are highly popular with users interested in IT topics (i.e. *GitHub* and *StackExchange*) and those users also tend to be quite active on Social Web platforms overall — it can be considered as an upper bound to the fraction of learners that are active on those five platforms and identifiable through robust and highly accurate means.

In Table 2.2 we split up the matches found according to the type of matching performed (Explicit, Direct or Fuzzy). On *Gravatar*, we relied exclusively on Explicit matching, while the vast majority of learners on *GitHub*
<table>
<thead>
<tr>
<th>MOOC</th>
<th>Year</th>
<th>#Learners</th>
<th>Gravatar</th>
<th>StackExchange</th>
<th>GitHub</th>
<th>LinkedIn</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solar Energy</td>
<td>2013</td>
<td>67,143</td>
<td>1,510</td>
<td>1,570</td>
<td>1,677</td>
<td>2,997</td>
<td>3,828</td>
</tr>
<tr>
<td>Solar Energy</td>
<td>2014</td>
<td>34,524</td>
<td>1,923</td>
<td>874</td>
<td>2,229</td>
<td>1,625</td>
<td>2,152</td>
</tr>
<tr>
<td>Solar Energy</td>
<td>2015</td>
<td>26,178</td>
<td>1,147</td>
<td>435</td>
<td>1,184</td>
<td>1,181</td>
<td>1,557</td>
</tr>
<tr>
<td>Introduction to Water Treatment</td>
<td>2014</td>
<td>10,458</td>
<td>457</td>
<td>129</td>
<td>430</td>
<td>427</td>
<td>1,548</td>
</tr>
<tr>
<td>Introduction to Water and Climate</td>
<td>2014</td>
<td>9,811</td>
<td>1,561</td>
<td>154</td>
<td>510</td>
<td>452</td>
<td>1,558</td>
</tr>
<tr>
<td>Technology for Biobased Products</td>
<td>2014</td>
<td>20,531</td>
<td>1,134</td>
<td>605</td>
<td>1,373</td>
<td>921</td>
<td>1,192</td>
</tr>
<tr>
<td>Next Generation Infrastructures¹</td>
<td>2014</td>
<td>38,682</td>
<td>9,987</td>
<td>8,477</td>
<td>16,220</td>
<td>4,274</td>
<td>6,801</td>
</tr>
<tr>
<td>Data Analysis²</td>
<td>2015</td>
<td>33,547</td>
<td>1,392</td>
<td>1,165</td>
<td>4,432</td>
<td>1,469</td>
<td>1,200</td>
</tr>
<tr>
<td>Pre-university Calculus³</td>
<td>2015</td>
<td>28,015</td>
<td>1,928</td>
<td>960</td>
<td>2,477</td>
<td>1,406</td>
<td>1,064</td>
</tr>
<tr>
<td>Introduction to Aeronautical Engine</td>
<td>2014</td>
<td>20,481</td>
<td>1,134</td>
<td>605</td>
<td>1,373</td>
<td>921</td>
<td>1,192</td>
</tr>
<tr>
<td>Introduction to Aeronautical Engine</td>
<td>2014</td>
<td>13,197</td>
<td>699</td>
<td>318</td>
<td>837</td>
<td>609</td>
<td>788</td>
</tr>
<tr>
<td>Topology in Condensed Matter⁴</td>
<td>2015</td>
<td>4,231</td>
<td>277</td>
<td>292</td>
<td>600</td>
<td>201</td>
<td>302</td>
</tr>
<tr>
<td>Framing⁵</td>
<td>2015</td>
<td>34,018</td>
<td>2,838</td>
<td>1,034</td>
<td>2,597</td>
<td>2,213</td>
<td>2,057</td>
</tr>
<tr>
<td>Solving Complex Problems⁶</td>
<td>2014</td>
<td>32,673</td>
<td>2,803</td>
<td>1,620</td>
<td>3,928</td>
<td>1,934</td>
<td>2,647</td>
</tr>
<tr>
<td>Delft Design Approach⁷</td>
<td>2014</td>
<td>13,543</td>
<td>1,319</td>
<td>514</td>
<td>1,376</td>
<td>1,085</td>
<td>1,124</td>
</tr>
<tr>
<td>Responsible Innovation⁸</td>
<td>2014</td>
<td>10,735</td>
<td>877</td>
<td>274</td>
<td>800</td>
<td>713</td>
<td>753</td>
</tr>
</tbody>
</table>

| Unique Users                       |      | 329,200   | 25,702   | 15,135        | 31,478 | 19,405   | 25,620  |

¹ Explores the challenges of global & local infra-structure (ICT, energy, water and transportation).
² Teaches the foundations of functional programming & how to apply them in practice.
³ Teaches data analysis skills using spreadsheets and data visualization.
⁴ Provides an overview of topological insulators, Majoranas, and other topological phenomena.
⁵ Analyzes how politicians debate and what the underlying patterns are framing and reframing.
⁶ How to solve complex problems with analytics based decision-making & solution designs.
⁷ How to design meaningful products & services.
⁸ How to deal with risks and ethical questions raised by the development of new technologies.

Table 2.1: Overview of the edx MOOCs under investigation, the number of learners registered to those MOOCs and the number of learners that could be matched (with either Explicit/Direct or Fuzzy matching) to our five Social Web platforms. Marked with ✯ (†) are those course/platform combinations where we were able to locate > 5% (> 10%) of the registered learners. The final row contains the unique number of users/learners (a learner may have taken several MOOCs) identified on each platform.

and StackExchange were also identified in this manner, with Direct and Fuzzy matching contributing little. On these platforms, users’ email addresses are either directly accessible (Gravatar and GitHub) or indirectly accessible (StackExchange provides the MD5 hash of its users’ email ad-
2.5. Results

In this section, we present an overview of our findings. As we collected different types of data (tweets vs. skills vs. source code) from different Social Web platforms, we describe the analysis conducted on each platform’s data traces independently in the following subsections.

2.5.1 Learners on Twitter

Our Twitter dataset consists of 25,620 unique users having written 12,314,067 tweets in more than 60 languages, which offers many insights into RQ 1.2. The majority language is English (68.3% of all tweets), followed by Spanish.

---

Note that StackExchange stopped the release of MD5 hashes in September 2013, thus we use the 2013 data dump for email matching and the September 2015 data dump for our content analysis.

---

To verify the quality of our matchings, for each platform, we sampled 50 users identified through any matching strategy and manually determined whether the correct linkage between the learner’s edX profile and the Social Web platform was found (based on the inspection of user profile information and content). We found our matching to be robust: of the 100 samples, we correctly linked 93 (StackExchange), 87 (GitHub), 97 (Twitter) and 95 (LinkedIn) respectively.

<table>
<thead>
<tr>
<th></th>
<th>Explicit</th>
<th>Direct</th>
<th>Fuzzy</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravatar</td>
<td>7.81%</td>
<td>—</td>
<td>—</td>
<td>7.81%</td>
</tr>
<tr>
<td>StackExchange</td>
<td>4.32%</td>
<td>0.01%</td>
<td>0.25%</td>
<td>4.58%</td>
</tr>
<tr>
<td>GitHub</td>
<td>9.04%</td>
<td>0.02%</td>
<td>1.23%</td>
<td>10.29%</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>—</td>
<td>0.48%</td>
<td>5.41%</td>
<td>5.89%</td>
</tr>
<tr>
<td>Twitter</td>
<td>—</td>
<td>0.67%</td>
<td>7.12%</td>
<td>7.78%</td>
</tr>
</tbody>
</table>

Table 2.2: Overview of the percentage of MOOC learners (329,200 overall) identified through the different matching strategies on the five selected Social Web platforms. A dash (—) indicates that for this specific platform/strategy combination, no matching was performed.

In contrast, the LinkedIn and Twitter platforms do not publish this type of user information and thus the majority of matches are fuzzy matches. Overall, the Direct approach has the least impact on the number of matches found.

Note that StackExchange stopped the release of MD5 hashes in September 2013, thus we use the 2013 data dump for email matching and the September 2015 data dump for our content analysis.
(7.3%), Dutch (3.1%), Portuguese (3.1%) and Russian (2.2%). The popularity of the Dutch language among our Twitter users can be explained by the fact that all MOOCs we consider in this analysis are offered by a Dutch university.

For each Twitter user with at least 100 English language tweets we estimated their age according to the approach described in Section 2.3.3. The results for our Twitter user set overall and three exemplary MOOCs (that is, we only consider users that participated in a particular MOOC) are shown in Figure 2.2: we have binned the estimations into six age brackets. The average MOOC learner is between 20 and 30 years of age, though we do observe that different types of courses attract slightly different audiences: In the Functional Programming MOOC, the 20-40 year old learners are overrepresented (compared to the “Overall” user set — computed across all eighteen MOOCs), while Framing and Responsible Innovation engage older learners to a larger than average degree.

We conduct an analogous analysis of our users’ gender distribution; the results are shown in Figure 2.3. The majority of MOOCs we investigate are anchored in engineering or the natural sciences, which traditionally attract a much larger percentage of male learners (in most parts of the world). This is reflected strongly in our Twitter sample: across all users with 100 or more English speaking tweets, 89% were identified as male. The MOOC with the highest skew in the distribution is Functional Programming with more than 96% of users identified as male. In contrast, the Framing and Robust Innovation exhibit the lowest amount of skewness: in both MOOCs, more than 20% of the users in our sample are classified as female.

The results we have presented provide us with confidence that microblog-based user profiling in the context of massive open online learning yields reliable outcomes. Future work will investigate the derivation of more complex and high-level attributes (such as personalities and learner type) from microblog data and their impact on online learning.

---

16 We generated these numbers based on Twitter’s language auto-detect feature.
17 Based on the ground truth data provided by 20,311 edX learners, the prediction precision is 36.5%.
18 The prediction precision is 78.3% based on the ground truth provided by 20,739 edX learners.
2.5. Results

Figure 2.2: Percentage of our Twitter users across eight age brackets. The “Overall” user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific.

Figure 2.3: Percentage of our Twitter users of each gender. The “Overall” user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific.

2.5.2 Learners on LinkedIn

LinkedIn user profiles are often publicly accessible, containing information about a user’s education, past and current jobs as well as their interests and skills. As shown in Table 2.1, for each of the MOOCs we were able to identify between 200 (Topology in Condensed Matter) and 2,997 (Solar Energy 2013) learners on the LinkedIn platform. To explore RQ 1.2 we focus on two types
of information in those profiles: job titles and skills. In our dataset, among the 19,405 collected LinkedIn profiles, 17,566 contain a job title (with on average 5.89 number of terms) and 16,934 contain one or more skills (37.42 skills on average).

Figure 2.4: Overview of the most frequent job title bigrams among the learners of the Data Analysis (top), Delft Design Approach (middle), and Responsible Innovation (bottom) MOOCs.
In Figure 2.5, exemplary for three MOOCs (Data Analysis, Responsible Innovation, and Delft Design Approach), we present the most frequently occurring bigrams among the job titles of our learners. Interestingly, the Data Analysis MOOC attracts a large number of self-proclaimed “software engineers” and “business analysts,” despite the fact that it covers elementary material (it is an introduction to spreadsheet-based data analysis & Python) which we consider users in this area to be already familiar with. In contrast, the Delft Design Approach and Responsible Innovation job title bigram distributions are more in line with our expectations — the most frequent bigrams are “project manager” and “co founder” respectively, positions for which knowledge about the risks and ethical questions of new technologies (Responsible Innovation) and the design of new products (Delft Design Approach) are very relevant to.

![Figure 2.5: Fraction of learners displaying n numbers of MOOC certificate.](image)

As prior works have indicated extrinsic factors such as recognition-by-others to play an important motivating role for MOOC learners, an explanation for the observed discrepancy between expected learners and actual MOOC participants, we also investigate to what extent our learners on LinkedIn present their MOOC achievements to the outside world. In Figure 2.5, we present a distribution of the number of MOOC certificates our users in the LinkedIn dataset list on their profile page. Each certificate represents a successfully completed MOOC. We limit our investigation to any certificate issued by the edX or Coursera platforms, as they offer a verifiable certificate interface to LinkedIn. We manually checked a random sample of 100 DelftX edX certificates listed by LinkedIn users to check whether each was actually issued to this specific user via edX. This was indeed the case for all sampled certificates. Overall, 9% of our users list one or more MOOC
certificates on their public profile with the majority of users (57%) having achieved one or two certificates only. A small fraction of learners (2%) is highly active in the MOOC learning community, having collected more than 15 certificates over time. Future work will investigate the impact of MOOC certificates on professional development through the lens of LinkedIn.

Lastly, we investigate to what extent the users’ listed skills on their LinkedIn profiles can be considered indicative of their course preferences (to enable course recommendations for instance). A user can list up to 50 skill on his profile — skills are not restricted to a pre-defined set, any keyword or short phrase can be added as a skill. Across all LinkedIn users in our dataset (19,405 users in total), the five most frequently mentioned skills are management (5,847 times), project management (4,894 times), java (4,087 times), microsoft office (4,073 times) and leadership (3,971 times). Thus, most of the users in our dataset present skills of themselves that are required for higher positions. We created a skill vocabulary by considering all skills mentioned at least once by a user in our dataset and then filtering out the fifty most frequent skills overall, leaving us with 28,816 unique skills. We create a user-skill matrix, where each cell represents the presence or absence of a skill in a user’s profile. We then applied truncated SVD \([52]\) to reduce the dimensions of the matrix to 50 and then employed t-SNE (described in Section 2.3.3) to visualize the structure of the data in a two dimensional space.

In Figure 2.6 we present the t-SNE based clustering of user skills exemplary for three pairs of MOOCs: Delft Design Approach vs. Topology of Condensed Matter, Data Analysis vs. Solar Energy 2015, and, Functional Programming vs. Framing. Recall, that a point in a plot represents a skill vector; t-SNE visually clusters data points together that are similar in the original (high-dimensional) skill space. The most distinct clustering can be observed for the final course pairing — users interested in functional programming are similar to each other, but different in their skill set from users interested in the analyses of political debates. This is a sensible result, which highlights the suitability of t-SNE for this type of data exploration. For the other two course pairings, the plots show less separation. In particular, for the Data Analysis vs. Solar Energy 2015 pairing, we observe a complete overlap between the two sets of users, i.e. there is no distinct set of skills that separates their interests. The pairing Delft Design Approach vs. Topology of Condensed Matter shows that the users of the design course have a larger spread of skills than those taking the physics MOOC. Still, the overlap in the skill set is considerable.
2.5. Results

Figure 2.6: t-SNE based visualization of LinkedIn skill vectors for pairs of MOOCs. Each data point represents one skill vector (i.e. one user).
2.5.3 Learners on StackExchange

Our StackExchange dataset consists of 86,672 questions (1% of all StackExchange questions posted), 197,504 answers (1.2% of all answers) and 418,633 comments, which were contributed by the 31,478 unique users we identified as MOOC learners among our courses. Given that 51.5% of the identified users registered for the Functional Programming MOOC, we focus our attention on the StackOverflow site within StackExchange (the Q&A site for programming-related questions), where our learners contributed 71,344 questions, 177,780 answers and 358,521 comments.

Driven by RQ 1.3, we first explored to what extent (if at all) MOOC learners change their question/answering behaviour during and after a MOOC. We restricted this analysis to the learners of the Functional Programming MOOC as those were by far the most active on StackOverflow. Among the 38,682 learners that registered for that MOOC, 8,068 could be matched to StackExchange. Of those users, 849 attempted to answer at least one question related to functional programming.

In Figure 2.7 (top) we plot month-by-month (starting in January 2014) the number of questions and answers by our learners that are tagged with “Haskell”, the functional language taught in the MOOC. Two observations can be made: (i) a subset of learners was already using Haskell before the start of the MOOC (which ran between 10/2014 and 12/2014), and, (ii) the number of Haskell questions posed by MOOC learners after the end of the MOOC decreased considerably (from an average of 32 questions per month before the MOOC to 19 per months afterwards), while the number of answers provided remained relatively stable. Figure 2.7 (bottom) shows that this trend is specific to the subset of MOOC learners: here we plot the frequency of “Haskell”-tagged questions and answers across all StackExchange users and observe no significant changes in the ratio between questions and answers. Finally, in Figure 2.7 (middle) we consider our learners’ uptake of functional programming in general, approximated by the frequency of questions and answers tagged with any of the nine major functional language names: Scala, Haskell, Common Lisp, Scheme, Coljure, Racket, Erlang, Ocaml, F#.

Finally, we also explored whether our MOOC learners have a similar expertise-dispensing behaviour as the general StackOverflow user population. To this end, we make use of the two expertise use types proposed in [169]: sparrows and owls. In short, sparrows are highly active users that...

---

19 Scala, Haskell, Common Lisp, Scheme, Coljure, Racket, Erlang, Ocaml, F#
2.5. Results

Figure 2.7: Overview of the number of StackOverflow questions and answers posted on a monthly basis between January 2014 and July 2015 by (i) our MOOC learners [top and middle], and (ii) all StackExchange users [bottom] for Haskell [top, bottom] and the nine major functional languages [middle]. Marked in gray is the time period of the Functional Programming MOOC. The dashed green line indicates the ratio of Questions to Answers in each month.

contribute a lot but do not necessarily increase the community’s knowledge. Their answers, while relevant, might be of low quality or low utility as they are motivated by reputation scores, and gamification elements of the platform. Owls on the other hand are users that are motivated to increase the overall knowledge contained in the platform. Owls are experts in the discussed topic, and they prove their expertise by providing useful answers to important and difficult questions. [169] proposed the mean expertise contribution (MEC) metric to capture measure expertise, based on answering quality, question debatableness and user activeness. Based on this metric, they determined 10.0% of the StackOverflow users to be owls. We derived MEC for our set of Functional Programming MOOC learners that are active on StackOverflow and found 21.0% of them to be owls. Thus, the
average MOOC learner is not only interested in gathering knowledge, but also in distributing knowledge to others, on a deeper level than the average StackExchange user.

2.5.4 Learners on GitHub

![GitHub Contributions Graph](image)

Figure 2.8: Month-by-month GitHub contributions in the Haskell language by the Functional Programming MOOC learners identified on GitHub.

Finally, with respect to RQ 1.3, we consider the concept of learning transfer, introduced in Section 2.3.3. As a social coding platform, GitHub is most suitable to explore programming-heavy MOOCs, thus we restrict our analysis (as in the previous section) to the Functional Programming MOOC. We are particularly interested in the extent of the learners’ functional programming after the end of the MOOC — our MOOC learners ask fewer topic-related questions (on StackExchange) over time, but does it also mean they program less in the language? To this end, we explored the 6,371,518 PushEvents we extracted from our MOOC learners between January 1, 2013 and June 30, 2015. Figure 2.8 provides a first answer to this question. The amount of Haskell programming by our learners was increasing slowly over time even before the start of the MOOC. A spike is visible in November 2014 (weeks 3-6 of the Functional Programming MOOC) and immediately after the end of the MOOC the contributions increase. However, by March 2015, i.e. three months after the end of the MOOC, the contributions are beginning to decline again towards nearly pre-MOOC levels.

In contrast to Haskell, we observe a sharp rise in “Scala” (the main functional language in industry) activities after the end of the MOOC which peak in November 2015. These functional activities are not evenly spread across all users though, only 32% of the users we identified on GitHub exhibited any type of functional language activities after the end of the Functional Language MOOC.
2.6 Conclusion

In the future, we will not only consider the addition of lines of codes in a particular language, but also perform fine-grained code analyses to investigate which specific concepts the learners picked up on in the MOOC and later employed in their own works.

2.6 Conclusion

In this chapter, we have provided a first exploratory analysis of learners’ Social Web traces across eighteen MOOCs and five globally popular Social Web platforms. We argue that MOOC-based learning analytics has much to gain from looking beyond the MOOC platform and accounting for the fact that learning events frequently happen beyond the immediate course environment. This study embraces the data traces learners leave on various Social Web platforms as integral parts of the distributed, connected, and open online learning ecosystem.

Focusing on RQ 1.1, we have found that on average 5% of learners can be identified on globally popular Social Web platforms. We observed a significant variance in the percentage of identified learners; in the most extreme positive case (Functional Programming/GitHub) we were able to match 42% of learners. We also found that learners with specific traits prefer different types of MOOCS (RQ 1.2) and we were able to present a first investigation into user behaviours (such as learning transfer over time) that are paramount in the push to make MOOCs more engaging and inclusive (RQ 1.3).

In this work we were only able to explore the possible contributions of each Social Web platform to enhance massive open online learning on a broad level. In future work, we will zoom in on each of the identified platforms and explore in greater detail how learners’ behaviours and activities can be explored to positively impact our understanding of massive open online learning and improve the learning experience.
Chapter 3

Learning Transfer

In this chapter, we follow Chapter 2 in which we have observed that over one-third of learners from a Functional Programming MOOC used GitHub to maintain their programming activities. While course completion is indeed an important measure of learning, we argue that another key measure is learning transfer: do learners actually use the newly acquired knowledge and skills to solve problems in practice? To answer the question, we combine the data traces from both edX and GitHub for analysis. The contributions of this chapter have been published in [30].
3.1 Introduction

The rising number of MOOCs enable people to learn & advance their knowledge and competencies in a wide range of fields. Learning, though, is only the first step; the application of the taught concepts is equally important, as knowledge that is learned but not frequently applied or activated is quickly unlearned \[166, 18, 148].

Existing investigations into student learning within MOOC environments are commonly based on pre- & post-course surveys and log traces generated within those environments by the individual learners [23]. While student learning is indeed an important measure of success, we argue that another key measure is the amount of learning transfer [94] that is taking place: do learners actually utilize the newly gained knowledge in practice? Are learners expanding their knowledge in the area over time or do they eventually move back to their pre-course knowledge levels and behaviours? These are important questions to address in the learning sciences, and their answers will enable us to shape the MOOCs of the future based on empirical evidence.

The main challenge researchers face in answering these questions is the lack of accessible, large-scale, relevant and longitudinal data traces outside of MOOC environments. While learners can be uniquely identified within a MOOC platform, at this point in time we have no general manner of capturing their behavioural traces outside of these boundaries.

Not all is lost though. Social Web platforms (Twitter being the prime example) have become a mainstay of the Web. They are used by hundreds of millions of users around the world and often provide open access to some — if not all — of the data generated within them. While most of these platforms are geared towards people’s private lives, in the past few years social Web platforms have also begun to enter our professional lives.

One such work-related social Web platform is GitHub\(^1\); it is one of the most popular social coding platforms world-wide with more than 10 million registered users. Hobbyists and professional programmers alike use GitHub to collaborate on programming projects, host their source code, and organize their programming activities. As GitHub was founded in 2007, we have potential access to log traces reaching several years into the past; moreover, its continuously increasing popularity will enable us to observe our learners over years to come. The potential of GitHub for behavioural mining has long been recognized by the software engineering research community where GitHub is

\(^1\text{https://github.com/}\)
3.1. Introduction

one of the most popular data sources to investigate how (groups of) people code.

Thus, for MOOCs with a strong focus on programming concepts, we consider GitHub to be one of the most detailed and openly accessible sources of learners’ relevant behavioral traces outside of the MOOC environment itself. Concretely, we analyze FP101x, an edx MOOC covering basic functional programming concepts. Of the 37,485 learners that registered for FP101x we matched 12,415 (33.1%) to their respective GitHub accounts, enabling a first large-scale analysis of the uptake of taught programming concepts in practice.

Here, we are foremost interested in exploring to what extent the course affects learners after it has ended. We are guided by the following three Research Questions:

RQ 2.1 To what extent do learners from a programming MOOC transfer the newly gained knowledge to practice?

RQ 2.2 What type of learners are most likely to make the transfer?

RQ 2.3 How does the transfer manifest itself over time?

Based on these guiding questions we have formulated seven research hypotheses which build on previous research efforts in workplace and classroom learning. In contrast to our work though, in these settings, the investigations are mostly based on questionnaires and interviews instead of behavioural traces. To the best of our knowledge, learning transfer has not yet been investigated in the context of MOOCs. Gaining deeper insights about the (lack of) learning transfer in MOOCs will lead to more informed discussions on the practical purposes and benefits of MOOCs. The main contributions of our work can be summarized as follows:

- We investigate to what extent learning transfer insights gained in workplace and classroom settings hold in the MOOC context. We find that the majority of findings are also applicable in the case of MOOCs.
- We introduce the use of external social Web-based data sources to complement learner traces within MOOC environments as a means to capture much more information about MOOC learners.

\[
2 \text{https://www.edx.org/course/introduction-functional-programming-delftx-fp101x}
\]
• We introduce GitHub as a specific large-scale data source to mine relevant longitudinal behavioural traces about learners before, during and after a programming-oriented MOOC.

3.2 Background

Meaningful, robust educational experiences transcend rote memorization of facts and leave the learner empowered to take on new problems and practice novel ways of thinking. In tracking student activity from the learning context (edX) to a real-world, practical one (GitHub) over a period of three years, the present study observes the first two of the three criteria of robust learning as outlined in [73]: (i) application in new situations different from the learning context, (ii) retained over the long-term, and (iii) prepares for future learning. Gaining a better understanding of how students apply what they learn in online learning environments over an extended time frame enables instructors to design future courses that induce more robust learning.

Some researchers [162, 158] have begun to look beyond traces generated in online learning environments, by utilizing post-course surveys or conducting post-course interviews with MOOC students.

Although the early studies of transfer stemmed from educational issues, the majority of recent learning transfer research literature is concerned with work-place training in Human Resource Development (HRD) [19]. With the recent influx of student activity data generated from digital learning environments, we can now empirically measure not only the rate of transfer, but other contributing factors as well. That, in tandem with the established surveying strategies used by HRD, promises to fundamentally change the way we think about measurable learning outcomes.

Learning transfer is the application of knowledge or skills gained in a learning environment to another context [10]. While training situations in professional environments have a clear target context (the job), this is not the case with most academic learning situations. Students are generally taught a broad set of skills and knowledge which they may apply in countless ways. This deliberately broad definition encapsulates both near transfer (to similar contexts) and far transfer (to dissimilar contexts) [13] and avoids the subjective question of how similar or different the learning context is from the target context, as we are only concerned with whether the student transferred the learned skills or knowledge beyond the learning context.
3.2. Background

Due to their rising popularity as a professional development tool and their roots as an educational resource, MOOCs serve as an ideal source of information to gain new insights on learning transfer. Studies have begun to discuss the learners’ intention to apply what they’ve learned in MOOCs but do not continue to track student activity beyond the learning platform [55]. The present research aims to reoperationalize [45] the understanding of learning transfer given the emerging possibilities of user modeling and learning analytics from the current standard of reported learning transfer towards observed learning transfer.

Yelon & Ford [173] offer a key distinction in transfer that differentiates open and closed skills. Open skill training programs include “leadership and interpersonal skills training,” and typical closed skill trainings include “various technical training and computer software training.” This emerges as an important distinction. In a study in which Blume et al. [19] found post-training knowledge (PTK) and post-training self-efficacy (PTSE) to have similar correlations with learning transfer, PTK and PTSE for closed skills resulted in lower correlation coefficients than for open skills. Independent of performance, self-efficacy is a person’s self-reported ability to successfully complete a future task [12]. Knowledge is measured as a result of a task—answering a quiz question correctly indicates possession of that knowledge [19].

Regarding the maintenance and persistence of learning transfer over time, Blume et al. [19] analyzed how the amount of time (the “lag”) between the end of training and the beginning of the transfer study affects learning transfer. They found that in studies with at least some lag time between training and testing, learners exhibited significantly lower post-training knowledge and post-training self-efficacy than those that tested students immediately following training.

In their survey of training professionals from 150 organizations, [138] report that 62% of employees in their organization “effectively apply what they learned in training” to their job immediately, 42% after six months, and 34% after one year. Other studies directly survey students in gathering self-reported data about learning transfer [92]. Another manner by which researchers have measured transfer is through assessment questions following instruction that, in order for students to answer correctly, would have to apply what they learned to a new context or problem [101, 2]. The present study examines transfer as a more naturally occurring, un-elicited phenomenon that the learners undertake and exhibit on their own accord.
3.3 FP101x

*Introduction to Functional Programming* (or short *FP101x*) is a MOOC offered on the *edX* platform. The course introduces learners to various functional programming concepts; all programming is performed in the functional language *Haskell*.

The first iteration of the course ran between October 15, 2014 and December 31, 2014. As is common in MOOCs today, learners were invited to participate in a pre-course and a post-course survey containing questions on the motivation of the learners, the perceived quality of the course, etc. In August 2015 we approached a subset of learners for an additional post-course survey.

The course was set up as an xMOOC: lecture videos were distributed throughout the 8 teaching weeks. Apart from lectures each week, exercises ("homeworks" and "labs") were distributed in the form of multiple choice (MC) questions. While homework questions evaluated learners on their understanding of high-level concepts and code snippets (e.g., "What is the result of executing [...]?") labs required learners to implement programs themselves. To enable fully automatic evaluations, all lab work was also assessed through MC questions. Each of the 288 MC questions was worth 1 point & could be attempted once. Answers were due 2 weeks after the release of the assignment. To pass the course, ≥ 60% of all MC questions had to be answered correctly.

Overall, 37,485 users registered for the course. Fewer than half (41%) engaged with the course, watching at least one lecture video. The completion rate was 5.25%, in line with similar MOOC offerings. Over 75% of the learners were male and more than 60% had at least a Bachelors degree.

3.4 Methodology

We first outline and justify the seven research hypotheses upon which we ground our work. Next, we describe in detail how to verify them empirically based on course questionnaire data, *edX* logs and *GitHub* data traces.

3.4.1 Research Hypotheses

Based on prior work we can make the following hypothesis associated with RQ 2.1:
3.4. Methodology

**H1** Only a small fraction of engaged learners is likely to exhibit learning transfer.

While previous works, e.g. [138], note transfer rates of up to 60%, we hypothesize our rate to be much lower, due to the natural setting we investigate, the difficulty of the topic (closed skills) and the generally low retention rate of MOOCs.

A large part of existing literature has focused on the different dimensions of a learner that may be indicative of a high or low transfer rate. Thus, the following research hypotheses are all related to RQ 2.2, which focuses on the type of learner exhibiting transfer.

**H2** Intrinsically motivated learners with mastery goals are more likely to exhibit learning transfer than extrinsically motivated learners.

[129] found that, in academic settings, mastery goals are more consistently linked to transfer success than performance goals. This was measured by instructors guiding students through either mastery- or performance-oriented experimental conditions and comparing their assessment scores. In line with intrinsic motivation, mastery goals are characterized by a learner’s intention to understand and develop new knowledge and abilities. Performance goals, extrinsically motivated, are those sought after in order to obtain positive judgements from others [139].

**H3** Learners expressing high self-efficacy are more likely to actively apply their trained tasks in new contexts.

In other words, in both academic and professional settings, if you believe that you are able to do something, you are more likely to try it [50, 78, 79, 125].

**H4** Experienced learners (high ability levels) are more likely to transfer trained skills and knowledge in order to maintain and improve performance levels [50].

**H5** Learners reporting a high personal capacity (time, energy and mental space) for transfer are more likely to actually exhibit learning transfer [78].

**H6** Learners exhibiting a high-spacing learning routine are more likely to exhibit learning transfer than learners with a low-spacing learning routine.
Here, high-spacing refers to a larger number of discrete learning sessions than low-spacing with few learning sessions each lasting a long time (i.e. “cramming”) [111, 112, 113].

Finally, for RQ 2.3 we investigate the following hypothesis:

H7 The amount of exhibited transfer decreases over time [138].

3.4.2 From Hypotheses To Measurements

Table 3.1 shows an overview of the data sources used to investigate each research hypothesis.

<table>
<thead>
<tr>
<th></th>
<th>Pre CS</th>
<th>Post CS</th>
<th>edX Logs</th>
<th>GitHub Logs</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H4</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H5</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H6</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>H7</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 3.1: Overview of the different data sources used to investigate each research hypothesis. CS refers to the conducted Course Surveys (before and after the course).

To explore H1 we relate learners’ performance during the course (as found in the edX logs) to their development activities on GitHub.

To determine the impact of learners’ motivation on learning transfer (H2), we distinguish learners based on their answers to several pre/post-course survey questions we manually established as being motivation-related. To determine intrinsic motivation we identified six question-answer pairs including the following two:

- What describes your interest for registering for this course?; Answer: My curiosity (in the topic) was the reason for me to sign up for this course [Pre CS, 5-point Likert]

\footnote{Due to space constraints, only a subset of the identified question/answer pairs are shown.}
3.4. Methodology

- Express your level of agreement with the following statement.; Answer: Course activities piqued my curiosity. [Post CS, 5-point Likert]

Similarly, for extrinsic motivation we determined nine appropriate question-answer pairs, including:

- What describes your interest for registering for this course? Choose the one that applies to you the most; Answer: My current occupation motivated me to enroll in the course. [Pre CS, 5-point Likert]

- Considering your experience in this, how much do you agree with the following statement?; Answer: The course was compulsory for me [Post CS, Multiple choice]

Learners’ belief in their ability to complete a task (H3), can be inferred based on a question asking the learners to express their level of agreement with a set of statements from the validated General Self-Efficacy Scale [29]:

- I can describe ways to test and apply the knowledge created in this course. [Post CS, 5-point Likert]

- I have developed solutions to course problems that can be applied in practice. [Post CS, 5-point Likert]

- I can apply the knowledge created in the course to my work or other non-class related activities. [Post CS, 5-point Likert]

The prior expertise of learners (H4) can both be inferred from survey questions as well as from the GitHub logs. The questions utilized are:

- Is your educational background related to (Functional) Programming? [Pre CS, 5-point Likert]

- Do you have professional experience in this field? [Pre CS, 5-point Likert]

The personal capacity (H5) of a learner is inferred based on two questions:

- Did any of the following negatively affect your participation in the course? [Post CS, 5-point Likert]

- Considering your experience in this course, how much did each of the technical issues affect your participation? [Post CS, 5-point Likert]

Responses to these questions allow learners to share which factors inhibited and distracted them from engaging with the course. Examples of responses to these questions range from personal problems, such as family obligations and medical issues, to technical trouble, such as slow Internet or hardware problems.
H6 considers the manner in which learners learn and can be inferred solely based on edX log traces which will be explained in more detail in the section below. Finally, H7, the extent to which functional programming is employed and applied by the learners over time can be inferred from GitHub logs alone.

**edX Logs**

For each learner, we collect all available traces (between October 1 and December 31, 2014), such as the learner’s clicks & views, provided answers to MC questions as well as forum interactions. Using the MOOCdb toolkit\[4\] we translate these low-level log traces into a data schema that is easily queryable.

To investigate H6, for each learner the learning routine is determined based on their edX logs. We partition the learners into low-spacing and high-spacing types following [111]. Initially, all learners are sorted in ascending order according to their total time on-site. Subsequently they are binned into ten equally-sized groups. Within each group, the learners are sorted according to the number of distinct sessions on the site and based on this ordering divided into two equally-sized subgroups: learners with few sessions (low-spacing) and learners with many sessions (high-spacing). In this manner, learners spending similar amounts of time (in total) on the course site can be compared with each other.

**GitHub Logs**

We identify edX learners on GitHub through the email identifiers attached to each edX and GitHub account. A third of all learners that registered to FP101x are also active on GitHub: 12,415 learners in total\[5\]. This is likely to be an underestimate of the true number of GitHub users (people generally have multiple email accounts), as we did not attempt to match accounts based on additional user profile information.

GitHub provides extensive access to data traces associated with public coding repositories, i.e. repositories visible to everyone\[6\]. GitHub is built

---

4\[http://moocdb.csail.mit.edu/\]  
5\[Note that the number is different from the one (16,220) we presented in Table 2.1 in Chapter 2. This is because: (i) we only consider learners registered before the end of the MOOC; (ii) Chapter 4 used GitHub Archive to match learners and we use GHTorrent here as it provides a more fine-grained record about users’ coding traces.\]  
6\[Data traces about private repositories are only available to the respective repository owner.\]
3.4. Methodology

around the git distributed revision control system, which enables efficient distributed and collaborative code development. GitHub not only provides relevant repository metadata (including information on how popular a repository is, how many developers collaborate, etc.), but also the actual code that was altered. As the GitHub Archive makes all historic GitHub data traces easily accessible, we relied on it for data collection and extracted all GitHub data traces available between January 1, 2013 and July 21, 2015. We then filtered out all traces that were not created by our edX learners, leaving us with traces from 10,944 learners. Of the more than 20 GitHub event types, we only consider the PushEvent as vital for our analysis.

```
{
  "_id" : ObjectId("55b6005de4b07ff432432dfe1"),
  "created_at" : "2013-03-03T18:36:09-08:00",
  "url" : "https://github.com/john/RMS/compare/1c55c4cb04...420e112334",
  "actor" : "john",
  "actor_attributes" : {
    "name" : "John Doe",
    "email" : "john@doe.com"
  },
  "repository" : {
    "id" : 2.37202e+06,
    "name" : "RMS",
    "forks" : 0,
    "open_issues" : 0,
    "created_at" : "2011-09-12T08:28:27-07:00",
    "master_branch" : "master"
  }
}
```

Figure 3.1: Excerpt of a GitHub PushEvent log trace.

Every time code is being updated ("pushed" to a repository), a PushEvent is triggered. Figure 3.1 contains an excerpt of the data contained in each PushEvent. The most important attributes of the event are the created_at timestamp (which allows us to classify events as before/during/after the running of FP101x), the actor (the user doing the "push") and the url, which contains the URL to the actual diff file. While the git protocol also allows a user to "push" changes by another user to a repository (which is not evident from inspecting the diff file alone), this is a rare occurrence among our learners: manually inspecting a random sample of 200 PushEvents showed 10 such cases. A diff file shows the difference between the last version of the repository and the new one (after the push) in terms of added and deleted code. An example excerpt is shown in Figure 3.2. For each of the identified 1,185,549 PushEvents by our learners, we crawled the corresponding diff

---

1. https://www.githubarchive.org/
Chapter 3. Learning Transfer

file, as they allow us to conduct a fine-grained code analysis. As a first step, we identified the additions and deletions a user conducts in each programming language based on the filename extensions found in the corresponding diff file. We consider code updates in the following nine functional languages as clear evidence for functional programming: Common Lisp, Scheme, Clojure, Racket, Erlang, Ocaml, Haskell, F# and Scala. We also log changes made in any of the other 20 most popular programming languages found on GitHub in the same manner. Any filename extension not recognized is first checked against a blacklist (which includes common filename extensions for images, compressed archives, audio files, etc.) and if not found, the change is classified as Other.

```
diff --git a/viewsA.rb b/viewsA.rb
index e37bca1..3ad75e4 100644
--- a/viewsA.rb
+++ b/viewsA.rb
@@ -26,6 +26,16 @@ def new
    @shift = Shift.new
end
...

diff --git a/config/routes.rb b/config/routes.rb
index e576929..27ce68f 100644
--- a/config/routes.rb
+++ b/config/routes.rb
@@ -29,6 +29,7 @@ put 'secondary'
    ...
```

Figure 3.2: Excerpt of a diff file. Two files were changed (viewsA.rb and routes.rb). The extension *.rb indicates code written in Ruby.

3.5 Results

We first present some basic characteristics of FP101x, before delving into the analyses of our research questions and hypotheses.

3.5.1 FP101x Overview

We partition our set of all registered FP101x learners according to two dimensions: (i) learners with and without a GitHub account, and (ii) learners with and without prior expertise in functional programming. In the latter case, we consider only those learners that could be identified on GitHub. We define Expert learners as those who used any of our nine identified functional programming languages before the start of the course to a meaningful degree.
3.5. Results

(i.e. more than 25 lines of functional code being added). The characteristics of these learner cohorts are listed in Tables 3.2 and 3.3.

When considering the GitHub vs. non-GitHub learners, we observe significant differences along the dimensions of engagement and knowledge:

- GitHub learners are on average more engaged with the course material (significantly more time spent on watching lecture videos and significantly more questions attempted).

- GitHub learners exhibit higher levels of knowledge (significantly more questions answered correctly).

Zooming in on the GitHub learners and their functional programming expertise, we find the differences to be enlarged: Expert learners have a higher completion rate (more than double that of non-Expert learners), attempt to solve significantly more problems and are significantly more accurate in answering. Experts are also more engaged in terms of forum usage - 8% post at least once compared to 4% of the non-Expert learners.

Finally, we note that we repeated this analysis on the subset of engaged learners only, where we consider all learners that attempted to solve at least one MC question or watched at least one video. While the absolute numbers vary, the trends we observe for the different partitions of learners in Tables 3.2 and 3.3 remain exactly the same.

3.5.2 Learning Transfer

Let us first consider the general uptake of functional programming languages. We can split each learner’s GitHub traces into three distinct sequences according to their timestamp: traces generated before, during and after FP101x. We are interested in comparing the before & after and will mostly ignore the activities generated during FP101x.

Expert Learners.

Overall, 1,721 of all GitHub learners have prior functional programming experience (our Expert learners). 1,165 of those are also engaged with FP101x (the remainder registered, but did not engage), leading to nearly a third (29.4%) of all engaged GitHub learners having pre-FP101x functional programming experience.

Most of our GitHub learners though are not continuously coding functionally: Figure 3.3 shows for each month of GitHub logs (January 2013 to
Chapter 3. Learning Transfer

<table>
<thead>
<tr>
<th></th>
<th>All Learners</th>
<th>GH Learners</th>
<th>Non-GH Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Enrolled learners</td>
<td>37,485</td>
<td>12,415</td>
<td>25,070</td>
</tr>
<tr>
<td>Completion rate</td>
<td>5.25%</td>
<td>7.71%</td>
<td>4.03%</td>
</tr>
<tr>
<td>% Learners who watched at least one video</td>
<td>40.84%</td>
<td>50.58%</td>
<td>36.02%</td>
</tr>
<tr>
<td>Avg. time watching video material (in min.) †</td>
<td>31.87</td>
<td>44.56</td>
<td>25.59</td>
</tr>
<tr>
<td>% Learners who tried at least one question</td>
<td>23.28%</td>
<td>31.94%</td>
<td>18.99%</td>
</tr>
<tr>
<td>Avg. # questions learners attempted to solve †</td>
<td>22.07</td>
<td>31.29</td>
<td>17.51</td>
</tr>
<tr>
<td>Avg. # questions answered correctly †</td>
<td>18.30</td>
<td>26.54</td>
<td>14.22</td>
</tr>
<tr>
<td>Avg. accuracy of learners’ answers †</td>
<td>16.36%</td>
<td>23.41%</td>
<td>12.86%</td>
</tr>
<tr>
<td>#Forum posts</td>
<td>8,157</td>
<td>3,726</td>
<td>4,431</td>
</tr>
<tr>
<td>% Learners who posted at least once</td>
<td>2.84%</td>
<td>4.27%</td>
<td>2.13%</td>
</tr>
<tr>
<td>Avg. # posts per learner †</td>
<td>0.22</td>
<td>0.30</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 3.2: Basic characteristics across all learners and their partitioning into GitHub (GH) and non-GitHub learners. Significant differences (according to Mann-Whitney) between GH and non-GH learners are marked with †(p < 0.001).

July 2015) the unique number of GitHub learners programming functionally - while in 2013 less than 250 of our GitHub learners were active per month, by 2015 this number has increased to nearly unique 600 active users a month. Thus, the trend to functional programming is generally increasing. Most learners though are not actively using functional languages on a monthly basis.

How much functional code do our engaged Expert learners produce over time? An answer to this question delivers Figure: for each month, the functional coding activities (calculated as the additions made in functional languages as a fraction of all additions made in recognized programming languages) are averaged across all engaged Expert learners. Again we observe that over the years functional programming has become more popular. By September 2014 (right before the start of FP101x), on average more than 36% of coding activities are functional. What is surprising (and somewhat counter-intuitive) is the steady decline of functional activities af-
3.5. Results

<table>
<thead>
<tr>
<th></th>
<th>Expert Learners</th>
<th>Non-Expert Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Enrolled learners</td>
<td>1,721</td>
<td>10,694</td>
</tr>
<tr>
<td>Completion rate</td>
<td>15.05%</td>
<td>6.53%</td>
</tr>
<tr>
<td>%Learners who watched at least one video</td>
<td>64.44%</td>
<td>48.35%</td>
</tr>
<tr>
<td>Avg. time watching video material (in min.)</td>
<td>69.61</td>
<td>40.53</td>
</tr>
<tr>
<td>%Learners who tried at least one question</td>
<td>48.69%</td>
<td>29.24%</td>
</tr>
<tr>
<td>Avg. #questions learners attempted to solve</td>
<td>57.86</td>
<td>27.02</td>
</tr>
<tr>
<td>Avg. #questions answered correctly</td>
<td>50.24</td>
<td>22.73</td>
</tr>
<tr>
<td>Avg. accuracy of learners’ answers</td>
<td>37.96%</td>
<td>21.06%</td>
</tr>
<tr>
<td>#Forum posts</td>
<td>1,612</td>
<td>2,114</td>
</tr>
<tr>
<td>%Learners who posted at least once</td>
<td>7.55%</td>
<td>3.74%</td>
</tr>
<tr>
<td>Avg. #posts per learners</td>
<td>0.94</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 3.3: Basic characteristics when partitioning the GitHub learners according to prior functional programming expertise. Significant differences (according to Mann-Whitney) between Expert and Non-Expert learners are marked with †(p < 0.001) and ‡(p < 0.01).

After the end of FP101x, if we restrict our engaged Expert Learners to those 542 learners with functional traces before and after FP101x (Figure 3.5), the results are more in line with our expectations: functional programming is continuously gaining in popularity and a peak in activities is observed in the two months following FP101x. Thus, 46.5% of engaged Expert learners did continue to program functionally after the end of FP101x.

Novice Learners.

Most interesting to use are the Novice Learners: to what extent do learners that did not program (meaningfully) in functional languages before FP101x take it up afterwards? We find 522 such learners — 4.3% of all GitHub learners. If we restrict ourselves to engaged GitHub learners, we are left with 336 Novice learners (8.5% of all engaged GitHub learners). Fig-

---

9The drop in July 2015 is explained by the non-complete log coverage of July (the log ends on July 21, 2015).
Figure 3.6 shows the evolution of their functional programming usage over time: the uptake after the end of *FP101x* is substantial, on average more than 35% of all activities are conducted subsequently in functional languages! While there is no substantial increase after the initial uptake over time, there is also no significant drop. Since the average can only provide limited insights, we drill down to the individual user level in Figure 3.7: the usage of functional programming is highly varied; 50% of the Novice Learners use it for less than 10% of their programming activities, while some learners almost exclusively code in functional languages. Finally, we also consider which functional languages these Novice learners code in. Figure 3.8 shows that a month after *FP101x* ended (January 2015), Haskell contributions made up 48% of all contributions, but continued dropping to a low of 14.5% in June 2015. Scala on the other hand (the most popular functional language in industrial settings) slowly rises in popularity over time and by June 2015 makes up roughly half of the functional contributions. Other functional languages play less of a role. Conducting a similar analysis on our engaged Expert learners (not shown here), we find that on average across all months, 47% ($\sigma = 7.4$) of all functional activities are in Scala, whereas 24.0% ($\sigma = 5.5$) are in Haskell. The distribution of functional languages is stable over time. The only outliers can be found in the three months of *FP101x*, where Haskell contributions rise significantly.

![Functional languages: Active users per month](image)

*Figure 3.3:* Number of unique users actively using a functional language. *FP101x* ran during the highlighted region.

Transfer Learning Hypotheses

On which learners should (or can) we investigate our seven research hypotheses? Ideally, we rely on all learners that engaged with the course and for whom GitHub traces are available. However, for Expert learners we are unable to determine the amount of transfer: since our analysis of functional
3.5. Results

Figure 3.4: Fraction of functional programming activities among the 1,165 engaged Expert Learners. FP101x ran during the highlighted region.

coding is based on activities in functional languages (instead of a more fine-grained analysis of the type of functional concepts employed), we are not able to determine whether learners that programmed functionally before acquired knowledge in FP101x and applied it in practice (a direction of future work). Only for the engaged Novice learners can we be confident that FP101x actually impacted their programming practice and that the observed transfer is likely a result of FP101x.

Considering H1, we observe a transfer rate of at least 8.5% (i.e. among the 3,965 engaged GitHub learners we found 336 Novice learners that began programming functionally after FP101x). This percentage can be considered as a lower bound, as we (due to the reasons listed above) do not consider engaged Expert learners here. Only a minority (70) of the 336 engaged Novice learners did pass FP101x, indicating that transfer and pass rate are related but not synonymous. In fact, while the 70 Novice learners that successfully completed the course remained mostly active until the final
course week (Figure 3.4), nearly 40% of all engaged Novice learners became inactive after week 1.

To investigate $H_2$, $H_3$, $H_4$, $H_5$ and $H_6$, for each hypothesis, we partition our 336 engaged Novice Learners who made the transfer according to the investigated dimensions (e.g. intrinsic vs. extrinsic motivation). Recall that the partitioning of the learners relies on their self-reported abilities in the pre- and post-course surveys. Similar to the retention rate, the return rate for such questionnaires is very low and many learners do not participate in these surveys for a variety of reasons. Table 3.4 shows the partitioning of our engaged Novice learners based on their survey data. The majority of learners cannot be assigned to a dimension due to a lack of data. Despite the low numbers, we do observe that the transfer learning hypotheses seem to hold in $FP101x$ (for those learners for which it is possible to measure their effect): learners are more likely to make the transfer if (i) they are intrinsically motivated, (ii) have high self-efficacy, (iii) are more experienced programmers, and (iv) report a high personal capacity. Even though the number of learners
3.5. Results

Figure 3.6: Fraction of functional programming activities among the 336 engaged Novice Learners with functional activities after FP101x. FP101x ran during the highlighted region.

we were able to investigate are small, we consider this as first evidence that transfer learning hypotheses also hold in the MOOC setting.

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>H2 Motivation</td>
<td></td>
</tr>
<tr>
<td>Extr.: 12</td>
<td>Intr.: 28</td>
</tr>
<tr>
<td>H3 Self-efficacy</td>
<td></td>
</tr>
<tr>
<td>High: 23</td>
<td>Low: 5</td>
</tr>
<tr>
<td>H4 Experience</td>
<td></td>
</tr>
<tr>
<td>A lot: 42</td>
<td>Little: 25</td>
</tr>
<tr>
<td>H5 Personal capacity</td>
<td>High: 22</td>
</tr>
</tbody>
</table>

Table 3.4: Partitioning of the 336 Novice learners according to several dimensions. The last column shows the number of learners that could not be assigned (N/A) to a dimension.

To answer H6 (high-spaced learners are more likely to transfer), we binned all GitHub learners according to their total time and number of distinct sessions in the FP101x edx environment, as outlined earlier. This creates 10 groups, with learners in Group 0 spending the least amount of time
Chapter 3. Learning Transfer

Figure 3.7: Distribution of functional programming activities among the 336 engaged Novice Learners with functional activities after FP101x. FP101x ran during the highlighted region.

Table 3.5: The number of Novice Learners falling into spacing groups.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Low spacing</th>
<th>High spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>30</td>
</tr>
</tbody>
</table>

and learners in Group 9 spending the most amount of time on the course site. Thus, each group contains those learners that roughly spent the same
3.5. Results

(amount of time on the site. Further, within each group, learners are divided according to the number of distinct sessions. In Table 3.5 we report how many engaged Novice learners fell into each group and which part of the group — the high-spacing or the low-spacing one. While 187 engaged Novice learners are classified as high-spacing, 149 are classified as low-spacing. Thus, there is some indication that H6 holds. However, the observed difference is rather small.

To conclude this section, we lastly consider H7. In contrast to the hypothesis (transfer decreases over time), we neither observe a significant decrease nor increase after the initial uptake as evident in Figures 3.6 and 3.7.

3.5.3 A Qualitative Analysis

We have found similarities and differences between transfer in classroom learning and our MOOC. Instead of speculating about the reasons for these differences, we designed a follow-up survey (containing 10 questions about learners’ functional programming experiences before and after FP101x) and distributed it to subsets of GitHub learners in August 2015. A second purpose of this questionnaire is to verify whether GitHub logs offer a good

---

10 All contacted learners had consented to additional contact.
approximation of our learners’ true behaviour. We partitioned the engaged GitHub learners into eight categories:

A **Novice** learners that **completed** the course but did **not transfer** (i.e. we did not observe functional GitHub traces after *FP101x*). #Survey responses: 131 (32% return rate).

B **Expert** learners that **completed** the course but did **not transfer**. #Survey responses: 15 (39%).

C **Novice** learners that **completed** the course and **transferred** (i.e. we observed functional GitHub traces after *FP101x*). #Survey responses: 11 (61%).

D **Novice** learners that **did not complete** the course, but **transferred**. #Survey responses: 1 (3%).

E **Expert** learners that **completed** the course and **continued programming functionally** (did they transfer?). #Survey responses: 20 (56%).

F **Expert** learners that did **not complete** the course but did **program functionally** after *FP101x*. #Survey responses: 8 (16%).
3.6 Conclusion

G Novice learners that were engaged in the course (but not completed) and did not transfer. #Survey responses: 93 (6%).

H Expert learners that were engaged in the course (but not completed) and did not transfer. #Survey responses: 4 (7% return rate).

How accurate are GitHub traces as approximation of learners’ functional programming activities? Of those learners we had identified as Novices, 63% also self-reported as such. Of the learners we estimated to have some prior functional programming experience, 77% self-reported prior experience. In particular, the latter number is intriguing: based on our stringent methodology, we can be confident that all of our identified Expert learners did indeed functionally program before FP101x, though about a quarter self-reports otherwise. Of the learners we identified as having demonstrated learning transfer, 88% also self-reported as doing so. Of those we identified as not having demonstrated learning transfer, only 37% self-reported of not having applied anything they had learnt. An explanation for this discrepancy is based on the non-exclusive use of GitHub: while 73% indicated that they use GitHub for either work or personal coding projects, 65% use a Private/Employer’s repository service, and 39% use BitBucket. While 73% is promising in that it accounts for nearly three quarters of all learners, we could only detect users who use the same email address for both their edX and GitHub account.

What are the main reasons for learners not to transfer their acquired functional programming skills? 80% of learners reporting a reason for not transferring their acquired skills report a lack of opportunities. Many learners go on to explain that the programming language standards in their work-place do not allow them to practice what they have learned. Another common sentiment is that it is difficult for some experienced programmers to suddenly change their ways. For example, when asked why they did not apply what they learned in FP101x to either work or personal projects, one respondent shared, “It takes time and effort to change old programming habits.” And another shared a similar sentiment: “[It’s] hard to think functionally after 25 years of imperative [programming] experience.”

3.6 Conclusion

We have investigated the extent of learning transfer in the MOOC setting and introduced the use of a social-Web based data source (i.e. GitHub) to complement the learner traces collected within MOOC environments. Focus-
ing on one-third of FP101x learners we were able to link to GitHub, we made several important findings:

(1) Most transfer learning findings from the classroom setting translate into the MOOC setup; large discrepancies were only found for H1: the amount of observed transfer and H7: the development of transfer over time.

(2) The observed transfer rate in MOOCs is low. We found that 8.5% of engaged learners were indeed exhibiting transfer to varying degrees in our GitHub traces. We acknowledge that a substantial amount of programming occurs outside of GitHub (e.g. in private employer repositories). While the traces we gather offer many new insights by following learners beyond the MOOC platform for an extended period of time, considering one external data source alone is a limiting factor.

(3) The amount of transfer, operationalized as the fraction of functional coding is varying highly: about 50% of the learners transferring code less than 10% of the time functionally, while a small minority almost exclusively turns to functional languages.

(4) After the end of FP101x, learners making the transfer quickly identified the most industrially-relevant functional language at this moment (Scala). Over time their activities in Scala increased significantly, while their activities in Haskell (the language of FP101x) decreased. Overall though, after the initial uptake of functional programming, the fraction of functional activities (between 35%-40%) of all coding activities remained constant.

The limitations of the current study (only 33% of learners could be coupled to a GitHub account and our exploratory analysis has been conducted on the programming language type level) naturally lead to three directions for future work: (i) instead of focusing on the amount of code added per language, a more detailed analysis will determine the particular functional concepts employed and match them with the course material, (ii) programming languages are taught in a variety of MOOCs, it is an open question whether the same methodology is applicable across a variety of courses, and lastly, (iii) we will move beyond the GitHub platform and consider alternative external data sources.
Chapter 4

Second Language Acquisition Modeling

In this chapter, we focus on investigating the problem of knowledge tracing in the setting of topic-specific MOOC platforms. Knowledge tracing, which uses computational algorithms to model learners’ mastery of knowledge being taught over time, is a well-established problem in computer-supported education. However, due to the lack of available datasets, this problem remains largely unexplored in the topic-specific MOOC platforms. With the three large-scale language learning datasets released by Duolingo [140], now we can gain a better understanding of learners in the topic-specific MOOC platforms. In particular, we investigate factors that are correlated with learners’ performance and then apply a machine learning technique to predict learners’ future performance. The contributions of this chapter have been published in [32].
4.1 Introduction

Knowledge tracing plays a crucial role in providing adaptive learning to learners [23]: by estimating a learner’s current knowledge state and predicting her performance in future interactions, learners can receive personalized learning materials (e.g., on the topics the learner is estimated to know the least about).

Over the years, various knowledge tracing techniques have been proposed and studied, including Bayesian Knowledge Tracing [31], Performance Factor Analysis [22], Learning Factors Analysis [27] and Deep Knowledge Tracing [35]. Notable is that most of the existing works focus on learning performance within mathematics in elementary school and high school due to the availability of sufficiently large datasets in this domain, e.g., ASSISTment and OLI [26, 65, 175, 86]. The generalization to other learning scenarios and domains remains under-explored.

Particularly, there are few studies attempted to explore knowledge tracing in the setting of Second Language Acquisition (SLA) [15]. Recent studies showed that SLA is becoming increasingly important in people’s daily lives and should gain more research attention to facilitate their learning process [17]. It remains an open question whether the existing knowledge tracing techniques can be directly applied to SLA modeling—the release of the Duolingo challenge datasets now enables us to investigate this very question.

Thus, our work is guided by the following research question: **RQ 3.1** What factors are correlated with learners’ language learning performance?

To answer the question, we first formulate six research hypotheses which are built on previous studies in SLA. We perform extensive analyses on the three SLA Duolingo datasets [30] to determine to what extent they hold. Subsequently, we engineer a set of 23 features informed by the analyses and use them as input for a state-of-the-art machine learning model, Gradient Tree Boosting [172, 51], to estimate the likelihood of whether a learner will correctly solve an exercise.

We contribute the following major findings: (i) learners who are heavily engaged with the learning platform are more likely to solve words correctly; (ii) contextual factors like the device being used and learning format are correlated with learners’ performance considerably; (iii) repetitive practice is a necessary step for learners towards mastery; (iv) Gradient Tree Boosting
is demonstrated to be an effective method for predicting learners’ future performance in SLA.

4.2 Data Analysis

Before describing the six hypotheses we ground our work in as well as their empirical validation, we first introduce the Duolingo datasets.

4.2.1 Data Description

To advance knowledge modeling in SLA, Duolingo released three datasets, collected from learners of English who already speak Spanish (EN-ES), learners of Spanish who already speak English (ES-EN), and learners of French who already speak English (FR-EN), respectively, over their first 30 days of language learning on the Duolingo platform. The task is to predict what mistakes a learner will make in the future. Table 4.1 shows basic statistics about each dataset. Interesting are in particular the last two rows of the table which indicate the unbalanced nature of the data: across all languages correctly solving an exercise is far more likely than incorrectly solving it. Note that the datasets contain rich information not only on learners, words and exercises but also on learners’ learning process, e.g., the amount of time a learner required to solve an exercise, the device being used to access the learning platform and the countries from which a learner accessed the Duolingo platform.

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>ES-EN</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Unique learners</td>
<td>1,213</td>
<td>2,643</td>
<td>2,593</td>
</tr>
<tr>
<td>#Unique words</td>
<td>2,178</td>
<td>2,915</td>
<td>2,226</td>
</tr>
<tr>
<td>#Exercises</td>
<td>326,792</td>
<td>731,896</td>
<td>824,012</td>
</tr>
<tr>
<td>#Words in all exercises</td>
<td>926,657</td>
<td>1,973,558</td>
<td>2,622,958</td>
</tr>
<tr>
<td>#Avg. words / exercise</td>
<td>2.84</td>
<td>2.7</td>
<td>3.18</td>
</tr>
<tr>
<td>%Correctly solved words</td>
<td>84%</td>
<td>86%</td>
<td>87%</td>
</tr>
<tr>
<td>%Incorrectly solved words</td>
<td>16%</td>
<td>14%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Table 4.1: Statistics of the datasets.

1 See http://sharedtask.duolingo.com/#task-definition-data
2 An exercise usually contains multiple words.
In our work, we use \textit{learning session} to denote the period from a learner’s login to the platform until the time she leaves the platform. We use \textit{learning type} to refer to the “session” information in the original released datasets, whose value can be \textit{lesson}, \textit{practice} or \textit{test}.

\subsection*{4.2.2 Research Hypotheses}

Grounded in prior works we explore the following hypotheses:

\textbf{H1} A learner’s \textit{living community} correlates with her language acquisition performance.

Previous works, e.g., [13] demonstrated that the surrounding living community is a non-negligible factor in SLA. For instance, a learner learning English whilst living in an English-speaking country is more likely to practice more often and thus more likely to achieve a higher learning gain than a learner not living in one.

\textbf{H2} The more \textit{engaged} a learner is, the more words she can master.

Educational studies, e.g., [24], have shown that a learner’s engagement can be regarded as a useful indicator to predict her learning gain, which is the number of mastered words in our case.

\textbf{H3} The \textit{more time} a learner spends on \textit{solving an exercise}, the more likely she will get it wrong.

\textbf{H4} Contextual factors such as the device being used (e.g. iOS or Android), learning type (lesson, practice or test) and exercise format (such as transcribing an utterance from scratch or formulating an answer by selecting from a set of candidate words) are correlated with a learner’s mastery of a word.

We hypothesize that, under specific contexts, a learner can achieve a higher learning gain due to the different difficulty level of exercises. For instance, compared to transcribing an utterance from scratch, a learner is likely to solve more exercises correctly when being provided with a small set of candidate words.

\textbf{H5} \textit{Repetition} is useful and necessary for a learner to master a word [17, 62, 45].

\textbf{H6} Learners with a high-spacing learning routine are more likely to learn more words than those with a low-spacing learning routine.

Here, high-spacing refers to a larger number of discrete learning sessions. Correspondingly, low-spacing refers to relatively few learning sessions, which
usually last a relatively long time. In other words, learners with a low-spacing routine tend to acquire words in a “cramming” manner [111, 49, 17].

4.2.3 Performance Metrics

We now define four metrics we use to measure a learner’s exercise performance.

**Learner-level Accuracy (Lear-Acc)** measures the overall accuracy of a learner across all completed exercises. It is calculated as the ratio between the number of words correctly solved by a learner and the total number of words she attempted.

**Exercise-level Accuracy (Exer-Acc)** measures to what extent a learner answers a particular exercise correctly. It is computed as the number of correctly solved words divided by the total number of words in the exercise.

**Word-level Accuracy (Word-Acc)** measures the percentage of times of a word being answered correctly by learners. For a word, it is calculated as the number of times learners provided correct answers divided by the total number of attempts.

**Mastered Words (Mast-Word)** measures how many words have been mastered by a learner. As suggested in [172], it takes about 17 exposures for a learner to learn a new word. Thus, we define a word being mastered by a learner only if (i) it has been exposed to the learner at least 17 times and (ii) the learner answered the word accurately in the remaining exposures.

4.2.4 From Hypotheses To Validation

To verify **H1**, we use the location (country) from where a learner accessed the Duolingo platform as an indicator of the learner’s living community. We first bin learners into groups according to their locations. Next, we calculate the average learner-level accuracy and the number of mastered words of learners in each group. We report the results in Table 4.2. Here we only consider locations with more than 50 learners. If a learner accessed the platform from more than one location, the learner would be assigned to all of the identified location groups. In contrast to our hypothesis, we do not observe the anticipated relationship between living community and language learning (e.g. Spanish-speaking English-learners living in the US do not perform better than other learners).
## Chapter 4. Second Language Acquisition Modeling

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Locations</th>
<th>Lear-Acc</th>
<th>Mast-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-EN</td>
<td>Avg.</td>
<td>83.57</td>
<td>3.37</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>84.12</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>83.01</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>83.66</td>
<td>3.46</td>
</tr>
<tr>
<td></td>
<td>AU</td>
<td>85.69</td>
<td>3.70</td>
</tr>
<tr>
<td>ES-EN</td>
<td>Avg.</td>
<td>85.91</td>
<td>2.74</td>
</tr>
<tr>
<td></td>
<td>CA</td>
<td>84.89</td>
<td>3.26</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>86.22</td>
<td>2.58</td>
</tr>
<tr>
<td></td>
<td>AU</td>
<td>85.82</td>
<td>3.50</td>
</tr>
<tr>
<td></td>
<td>GB</td>
<td>83.94 *</td>
<td>3.30</td>
</tr>
<tr>
<td></td>
<td>NL</td>
<td>87.15</td>
<td>2.86</td>
</tr>
<tr>
<td>EN-ES</td>
<td>Avg.</td>
<td>87.62</td>
<td>4.39</td>
</tr>
<tr>
<td></td>
<td>CO</td>
<td>87.49</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>US</td>
<td>87.98</td>
<td>5.02</td>
</tr>
<tr>
<td></td>
<td>ES</td>
<td>87.85 *</td>
<td>5.66 *</td>
</tr>
<tr>
<td></td>
<td>MX</td>
<td>86.92 *</td>
<td>3.71 *</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>88.95</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>DO</td>
<td>87.26</td>
<td>4.40</td>
</tr>
<tr>
<td></td>
<td>AR</td>
<td>89.58</td>
<td>4.75</td>
</tr>
<tr>
<td></td>
<td>VE</td>
<td>89.47 *</td>
<td>4.99</td>
</tr>
<tr>
<td></td>
<td>PE</td>
<td>88.83</td>
<td>4.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lear-Acc</th>
<th>Mast-Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-EN</td>
<td>ES-EN</td>
</tr>
<tr>
<td># Exercises Attempted</td>
<td>-0.05 *</td>
</tr>
<tr>
<td># Words Attempted</td>
<td>-0.06 *</td>
</tr>
<tr>
<td>Time Spent</td>
<td>-0.15 *</td>
</tr>
</tbody>
</table>

### Table 4.2: Avg. learner-level accuracy (%) and the number of mastered words of learners living in different locations (approximated by the countries from which learners have finished the exercises). Significant differences (compared to Avg., according to Mann-Whitney) are marked with * ($p < 0.001$).

### Table 4.3: Pearson Correlation between learner engagement (measured by # attempted exercises/words and the amount of time spent in learning) and learner-level accuracy as well as # mastered words. Significant differences are marked with * ($p < 0.001$).

For H2 (learner engagement), we consider three ways to measure engagement with the platform: (i) number of attempted exercises, (ii) number of...
4.2. Data Analysis

attempted words and (iii) amount of time spent learning. To quantify the relationship between learners’ engagement and their learning gain, we report the Pearson correlation coefficient between the three engagement metrics and Lear-Acc as well as Mast-Word (Table 4.3). We note a consistent negative correlation between accuracy and our engagement metrics. This is not surprising, as more engagement also means more exposure to novel vocabulary items. When examining the number of mastered words, we can conclude that—as stated in $H_2$—higher engagement does indeed lead to a higher learning gain. This motivates us to design engagement related features for knowledge tracing models.

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>ES-EN</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>-0.16 *</td>
<td>-0.18 *</td>
<td>-0.18 *</td>
</tr>
</tbody>
</table>

Table 4.4: Pearson Correlation between the amount of time spent in solving each exercise and exercise-level accuracy. Significant differences are marked with * ($p < 0.001$).

To determine the validity of $H_3$, in Table 4.4 we report the Pearson correlation coefficient between the amount of time spent in solving each exercise and the corresponding exercise-level accuracy. The moderate negative correlation values indicate that the hypothesis holds to some extent.

For $H_4$, we investigate three types of contextual factors: (i) device used (i.e., Web, iOS, Android); (ii) learning type (i.e., Lesson, Practice, Test) and (iii) exercise format (i.e., Reverse Translate, Listen, Reverse Tap). To verify whether these contextual factors are correlated with learners’ exercise performance, we partition exercises into different groups according to the contextual condition in which they were completed and calculate the average of their exercise-level accuracy within each group. Table 4.5 shows the results. Interestingly, learners with iOS devices perform better than those using Web or Android. Learners’ learning accuracy is highest in the Lesson type. Learning formats are also likely to have a positive effect: Reverse Tap achieves the highest accuracy followed by Reverse Translate and then Listen. This result is not surprising as active recall of words is more difficult than recognition. Finally, we note for English learners who speak Spanish (EN-ES) and Spanish learners who speak English (ES-EN), the accuracy of Reverse Translate is considerably higher than Listen, which is not the case in FR-EN (where both are comparable). These results suggest that contextual factors should be taken into account in SLA modeling.
Table 4.5: Average exercise-level accuracy (%) in different contextual conditions. Significant differences (compared to Avg., according to Mann-Whitney) are marked with *(p < 0.001).

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>ES-EN</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>84.29</td>
<td>86.31</td>
<td>87.96</td>
</tr>
<tr>
<td>Client</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Web            | 80.64  | 85.44  | 85.68  | *
| iOS            | 86.45  | 87.90  | 88.10  | *  
| Android        | 83.92  | 84.88  | 88.92  | *  
| Session        |        |        |        |
| Lesson         | 85.43  | 87.23  | 88.76  | *  
| Practice       | 80.94  | 83.92  | 84.19  | *  
| Test           | 82.19  | 84.34  | 84.66  | *  
| Format         |        |        |        |
| Reverse Translate | 77.92  | 85.88  | 85.42  | *  
| Listen         | 78.30  | 77.01  | 82.78  | *  
| Reverse Tap    | 92.51  | 94.84  | 95.48  | *  

We investigate H5 from two angles. Firstly, we investigate whether words with very different exposure amounts will differ from each other in terms of word-level accuracy as they are practiced by learners to different degrees. For this purpose, we only retain words with more than \( n \) exposures (with \( n \) being \( \geq 1, \geq 10, \geq 20, \geq 50, \geq 100 \)) and calculate Pearson correlation coefficient between the word-level accuracy and their number of exposures (Table 4.6). As expected, the more low-exposure words we filter out, the higher the average word-level accuracy and the stronger the correlation scores (albeit at best these are moderate correlations).

Secondly, we believe that whether a learner will solve a word correctly (0 mean solving correctly and 1 incorrectly) is correlated with two factors that are related to word repetition. One factor is the number of previous attempts that a learner has for a word, and the other is the amount of time elapsed since her last attempt at the word. Therefore, we compute Pearson correlation coefficient between learners’ performance on exercises and the two repetition related factors (Table 4.7). The resulting correlations are even weaker than in our preceding analysis, though they do point towards a (very) weak relationship: if a learner gets more exposed to a word or practices the
## Data Analysis

### Table 4.6: Avg. word-level accuracy (%) of words with different number of exposures.

<table>
<thead>
<tr>
<th># Previous attempts</th>
<th>Word-Acc</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥ 1</td>
<td>2,178</td>
<td>72.30 -0.08 *</td>
</tr>
<tr>
<td>≥ 10</td>
<td>1,007</td>
<td>75.01 0.13 *</td>
</tr>
<tr>
<td>≥ 20</td>
<td>756</td>
<td>75.78 0.15 *</td>
</tr>
<tr>
<td>≥ 50</td>
<td>756</td>
<td>76.41 0.19 *</td>
</tr>
<tr>
<td>≥ 100</td>
<td>580</td>
<td>77.47 0.25 *</td>
</tr>
</tbody>
</table>

### Table 4.7: Pearson Correlation between learner performance and the number of previous attempts and the amount of time elapsed since the last attempt for a word.

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>ES-EN</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td># Previous attempts</td>
<td>-0.05 *</td>
<td>-0.04 *</td>
<td>-0.07 *</td>
</tr>
<tr>
<td>Time elapsed</td>
<td>0.05 *</td>
<td>0.06 *</td>
<td>0.07 *</td>
</tr>
</tbody>
</table>

Lastly, to study H6, we partition all learners into low-spacing and high-spacing groups according to [111]. Initially, all learners are sorted in ascending order according to their total time spent in learning words. Subsequently, these learners are binned into ten equally-sized groups labeled from 0 (spending the least amount of time) to 9 (spending the most amount of time). Therefore, we can regard learners from the same group as learning roughly the same amount of time. Next, within each group, the learners word more frequently, she is more likely to get it correct. Clearly, the results indicate that other factors at play here too.
Figure 4.1: The average learner-level accuracy, i.e., Lear-Acc (Top), and the average number of mastered words, i.e., Mast-Word (Bottom), of learners in high-spacing and low-spacing groups.
are sorted based on their number of distinct learning sessions\textsuperscript{3}, and we further divide them into two equally-sized subgroups: learners with few sessions (low-spacing) and learners with many sessions (high-spacing). In this way, learners spending similar total amounts of time can be compared with each other. We plot the average learner-level accuracy as well as the number of mastered words within each low-spacing and high-spacing subgroup in Figure 4.1. We do not observe consistent differences between low-spacing and high-spacing groups. Therefore, we conclude H6 to not hold.

4.3 Knowledge Tracing Model

We now describe the machine learning model we adopt for knowledge tracing and then introduce our features.

4.3.1 Gradient Tree Boosting

Various approaches have been proposed for modeling learner learning. Two representatives are Bayesian Knowledge tracing \cite{10} and Performance Factor Analysis \cite{122}, both of which have been studied for years. Inspired by the recent wave of deep learning research in different domains, deep neural nets were also recently applied to track the knowledge state of learners \cite{125,168,175}. In principal, all of these methods can be adapted to predict learners’ performance in SLA. As our major goal is to investigate the usefulness of the designed features, we selected a robust model that is able to take various types of features as input and works well with skewed data. Gradient Tree Boosting (GTB) is a machine learning technique which can be used for both regression and classification problems \cite{172}. It is currently one of the most robust machine learning approaches that is employed for a wide range of problems \cite{33}. It can deal with various types of feature data and has reliable predictive power when dealing with unbalanced data (as in our case). We selected it over a deep learning approach as we aim to build an interpretable model.

4.3.2 Feature Engineering

Based on the results in Section 4.2, we designed 23 features. The features are categorized into two groups: features directly available in the datasets (7

\textsuperscript{3}Here we consider all learning activities occurring within 60 minutes as belonging to the same learning session.
Chapter 4. Second Language Acquisition Modeling

given features) and features derived from the datasets (16 derived features).
Note that the features differ in their granularity—they are computed per
learner, or per word, per exercise or a combination of them, as summarized
in Table 4.8.

<table>
<thead>
<tr>
<th>Features</th>
<th>Granularity Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User</td>
</tr>
<tr>
<td>Learner ID</td>
<td>✓</td>
</tr>
<tr>
<td>Word</td>
<td>✓</td>
</tr>
<tr>
<td>Countries</td>
<td>✓</td>
</tr>
<tr>
<td>Format</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
<tr>
<td>Device</td>
<td></td>
</tr>
<tr>
<td>Time spent (exercise)</td>
<td></td>
</tr>
<tr>
<td># Exercises attempted</td>
<td>✓</td>
</tr>
<tr>
<td># Words attempted</td>
<td>✓</td>
</tr>
<tr>
<td># Unique words attempted</td>
<td>✓</td>
</tr>
<tr>
<td># sessions</td>
<td>✓</td>
</tr>
<tr>
<td>Time spent (learning)</td>
<td>✓</td>
</tr>
<tr>
<td># Previous attempts</td>
<td>✓</td>
</tr>
<tr>
<td># Correct times</td>
<td>✓</td>
</tr>
<tr>
<td># Incorrect times</td>
<td>✓</td>
</tr>
<tr>
<td>Time elapsed</td>
<td>✓</td>
</tr>
<tr>
<td>Word-Acc</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (exercise)</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (word)</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (session)</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (word-session)</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (word-correct)</td>
<td>✓</td>
</tr>
<tr>
<td>Std. timestamps (word-incorrect)</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 4.8: Granularity levels on which each feature is retrieved or computed. Features marked with b are used as input in the baseline provided by the benchmark organizers.

Given features:

- **Learner ID**^b^: the 8-digit, anonymized, unique string for each learner;

- **Word**^b^: the word to be learnt by a learner;
4.3. Knowledge Tracing Model

- **Countries**: a vector of dimension N (N denotes the total number of countries) with binary values indicating whether a learner complete an exercise in one or multiple countries;

- **Format**: the exercise format in which a learner completed an exercise, i.e., Reverse Translate, Reverse Tap and Listen;

- **Type**: the learning type in which a learner completed an exercise, i.e., Lesson, Practice and Test;

- **Device**: the device platform which is used by a learner to complete an exercise, i.e., iOS, Web and Android;

- **Time spent (exercise)**: the amount of time a learner spent in solving an exercise, measured in seconds;

Derived features:

- **# Exercises attempted**: the number of exercises that a learner has attempted in the past;

- **# Words attempted**: the number of words that a learner has attempted in the past;

- **# Unique Words attempted**: the number of unique words a learner has attempted in the past;

- **# Sessions**: the number of learning sessions a learner completed;

- **Time spent (learning)**: the total amount of time a learner spent learning, measured in minutes;

- **# Previous attempts**: a learner’s number of previous attempts at a specific word;

- **# Correct times**: the number of times that a learner correctly solved a word;

- **# Incorrect times**: the number of times that a learner incorrectly solved a word;

- **Time elapsed**: the amount of time that elapsed since the last exposure of a word to a learner;

- **Word-Acc**: the word-level accuracy that a learner gained for a word in the training dataset;
Chapter 4. Second Language Acquisition Modeling

- **Std. timestamps (exercise):** the standard deviation of the timestamps that a learner solved exercises;

- **Std. timestamps (word):** the standard deviation of the timestamps that a learner solved a word;

- **Std. timestamps (session):** the standard deviation of timestamps that a learner logged in to start a learning session;

- **Std. timestamps (word-session):** the standard deviation of session starting timestamps that a learner solved a word;

- **Std. timestamps (word-correct):** the standard deviation of timestamps that a learner answered a word correctly;

- **Std. timestamps (word-incorrect):** the standard deviation of timestamps that a learner answered a word incorrectly.

Finally, we note that none of the features in our feature set make use of external data sources. We leave the inclusion of additional data sources to future work.

### 4.4 Experiments

In this section, we first describe our experimental setup and then present the results.

#### 4.4.1 Experimental Setup

Each of the three Duolingo datasets consists of three parts: TRAIN and DEV sets for offline experimentations and one TEST set for the final evaluation. We use the TRAIN and DEV sets to explore features that are useful in predicting a learner’s exercise performance and then combine TRAIN and DEV sets to train the GTB model; we report the model’s performance on the TEST set.

We trained the GTB model using XGBoost, a scalable machine learning system for tree boosting [34]. All model parameters were optimized through grid search and are reported in Table 4.9.

---

[34] For a detailed explanation of the parameters, please refer to https://github.com/dmlc/xgboost/blob/v0.71/doc/parameter.md.
4.4. Experiments

We also report the official baseline provided by the benchmark organizers as comparison. The baseline is a logistic regression model which takes six features as input, which include learner ID, word, format and three morphosyntactic features of the word (e.g., Part of Speech). As suggested by the benchmark organizers, we use the AUC and F1 scores as our evaluation metrics.

<table>
<thead>
<tr>
<th></th>
<th>FR-EN</th>
<th>ES-EN</th>
<th>EN-ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning_rate</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>n_estimators</td>
<td>800</td>
<td>1100</td>
<td>1550</td>
</tr>
<tr>
<td>max_depth</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>min_child_weight</td>
<td>7</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>gamma</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>subsample</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>colsample_bytree</td>
<td>0.7</td>
<td>0.7</td>
<td>0.85</td>
</tr>
<tr>
<td>reg_alpha</td>
<td>4</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 4.9: Model parameters of the GTB model; determined by using grid search per dataset.

4.4.2 Results

In order to investigate the features described in Section 4.3.2, we report in Table 4.10 different versions of GTB training, starting with three features (Learner ID, Word, Format) and adding additional features one at a time. We incrementally added features according to the order presented in Section 4.3.2 and only kept features that boost the prediction performance (i.e. the AUC score improves on the DEV set). Among all 23 evaluated features, seven are thus useful for SLA modeling. Here, we only report the results in the ES-EN dataset; we make similar observations in the other two datasets. In contrast to our expectations, a large number of the designed features did not boost the prediction accuracy. This implies that further analyses of the data and further feature engineering efforts are necessary. The extraction of features from external data sources (which may provide insights in the difficulty of words, the relationship between language families and so on) is also left for future work.

In our final prediction for the TEST set, we combine the TRAIN and DEV data to train the GTB model with the nine features listed in Table 4.10 and learner ID as well as the word as input. The results are shown in Table 4.11.


<table>
<thead>
<tr>
<th>Feature</th>
<th>TRAIN</th>
<th>DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner ID &amp; Word &amp; Format</td>
<td>0.8095</td>
<td>0.7758</td>
</tr>
<tr>
<td>Mode</td>
<td>0.8111</td>
<td>0.7780</td>
</tr>
<tr>
<td>Client</td>
<td>0.8137</td>
<td>0.7790</td>
</tr>
<tr>
<td>Time spent (exercise)</td>
<td>0.8270</td>
<td>0.7828</td>
</tr>
<tr>
<td># Previous attempts</td>
<td>0.8323</td>
<td>0.7835</td>
</tr>
<tr>
<td># Wrong times</td>
<td>0.8348</td>
<td>0.7871</td>
</tr>
<tr>
<td>Std. time (word-session)</td>
<td>0.8348</td>
<td>0.7871</td>
</tr>
</tbody>
</table>

Table 4.10: Experimental results reported in AUC on ES-EN. Each row indicates a feature added to the GBT feature space; the model of row 1 has three features.

Compared to the logistic regression baseline, GBT is more effective with a 6% improvement in AUC and 83% improvement in F1 on average.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AUC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-EN</td>
<td>0.7707</td>
<td>0.2814</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.8153 *</td>
<td>0.4145 *</td>
</tr>
<tr>
<td>GTB</td>
<td>0.8013 *</td>
<td>0.3436 *</td>
</tr>
<tr>
<td>ES-EN</td>
<td>0.7456</td>
<td>0.1753</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.8013 *</td>
<td>0.3436 *</td>
</tr>
<tr>
<td>GTB</td>
<td>0.8210 *</td>
<td>0.3889 *</td>
</tr>
<tr>
<td>EN-ES</td>
<td>0.7737</td>
<td>0.1899</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.8210 *</td>
<td>0.3889 *</td>
</tr>
</tbody>
</table>

Table 4.11: Final prediction results on the TEST data. Significant differences (compared to Baseline, according to paired t-test) are marked with * (p < 0.001).

4.5 Conclusion

Knowledge tracing is a vital element in personalized and adaptive educational systems. In order to investigate the peculiarities of SLA and explore the applicability of existing knowledge tracing techniques for SLA modeling, we conducted extensive data analyses on three newly released Duolingo datasets. We identified a number of factors relating to learners’ learning performance in SLA. We extracted a set of 23 features from learner trace data and used them as input for the GTB model to predict learners’ knowledge state. Our experimental results showed that (i) a learner’s engagement plays an important role in achieving good exercise performance; (ii) contextual factors like the device being used and learning format should be taken into
account for SLA modeling; (iii) repetitive practice of words and exercises are related to learners’ performance considerably; (iv) GTB can effectively use some of the designed features for SLA modeling and there is a need for further investigation on feature engineering. Apart from the future work already outlined in previous sections, we also plan to investigate deep knowledge tracing approaches and the inclusion of some of our rich features into deep models, inspired by [175]. Also, instead of developing a one-size-fits-all prediction model, it will be interesting to explore subsets of learners that behave similarly and develop customized models for different learner groups.
Chapter 5

Enabling MOOC Learners to Solve Real-world Paid Tasks

In this chapter, we focus on investigating whether learners can apply the knowledge acquired from a MOOC to solve real-world tasks, e.g., freelancing tasks collected from online marketplaces like Upwork or witmart\(^1\), which can be solved with the knowledge taught in the MOOC. If learners are capable of solving such tasks, it becomes possible that learners can learn with a MOOC and apply the newly acquired knowledge to earn money at the same time. Ultimately, we envision a recommender system that automatically retrieves paid tasks relevant to a MOOC from online marketplaces and presents these tasks to learners to solve, as a possible means to help learners, who do not have a large amount of time for learning because of the need to work and earn a living, to benefit from MOOCs. To investigate the potential of the proposed vision, we consider the specific case of Data Analysis: Take It to the MAX() (a MOOC teaching data analysis in edx). We manually select a set of relevant tasks from Upwork and offer them to learners in the MOOC as bonus exercises to solve. Based on our experimental design, we also investigate the impact of real-world tasks on the MOOC learners. The contributions of this chapter have been published in [28].

\(^1\)http://www.witmart.com
5.1 Introduction

In 2011, the first MOOCs started out with the promise of educating the world. To this day, this promise remains largely unfulfilled, as MOOCs struggle with student engagement and retention rates — on average, only 6.5% of MOOC learners complete a course and those who do often already have a higher degree [22]. At the same time though, the potential reach of MOOCs was visible from the very beginning: learners from 162 different countries engaged with the very first MOOC (Circuits and Electronics) offered on the edX platform [21].

Among the many reasons for learners’ disengagement from a course are also financial ones: learning is superseded by the need to work and earn a living. Our ultimate vision is to pay learners to take a MOOC, thus enabling learners from all financial backgrounds to educate themselves. But how can we achieve this at scale? We believe that online work platforms such as Upwork and witmart can be an important part of the solution; if we were able to automatically recommend paid online work tasks to MOOC learners which are related and relevant to the MOOC content, the financial incentive would enable more learners to remain engaged in the MOOC and continue learning.

Figure 5.1 shows a high-level overview of our vision: online work task platforms are continuously monitored for newly published work tasks; a recommender system maintains an up-to-date course model of every ongoing MOOC and determines how suitable each work task is for every ongoing course and course week. At any given moment, the suitable open work tasks are shown alongside the course material on the MOOC platform, together with the possible financial gain and their level of difficulty.

While we do not claim this vision as the solution for MOOCs to single-handedly “lift ... people out of poverty,” [57], we strongly believe this to be a step in the right direction and something to build upon.

To lay the groundwork, we investigate the feasibility of letting MOOC students solve real world tasks from an online work market place. In a pilot study presented here, we manually selected a number of paid tasks from Upwork and offered them to learners of the EX101x MOOC (Data Analysis: Take It to the MAX(), offered on edX) as bonus exercises. We illustrate that it is indeed feasible to expect students to be able to earn money while taking a MOOC.
5.1. Introduction

Based on these encouraging initial results we then expand our investigation and analyse the realm of online work platforms and their suitability for our vision along a number of dimensions including payments, topical coverage and task time.

Lastly, it is worth nothing that our experimental setup not only allows us to investigate learning enabling methods (i.e. paying learners), but also learner motivations: we expect that real-world tasks (as shown in the bonus exercises) engage learners more than artificially created course tasks.

The work we present in this chapter is guided by the following four Research Questions:

**RQ 4.1** Are MOOC learners able to solve real-world (paid) tasks from an online work platform with sufficient accuracy and quality?

**RQ 4.2** How applicable is the knowledge gained from MOOCs to paid tasks offered by online work platforms?

**RQ 4.3** To what extent can an online work platform support MOOC learners (i.e., are there enough tasks available for everyone)?

**RQ 4.4** What role do real-world (paid) tasks play in the engagement of MOOC learners?
By answering these questions, we expect to provide solid evidence to the feasibility of the proposed design, i.e., automatically retrieving relevant tasks from online marketplaces and recommend them to learners, we expect to financially enhance MOOC learners and help them achieve professional development in the long run.

5.2 Background

This study represents a movement towards MOOCs truly living up to their name with respect to their openness. The current demographic of MOOC participants is predominantly educated males from developed countries [30, 43, 73, 132, 90]. Simply putting the content out there on the Web may not be enough to justify calling it “open”. Although it is available, it is not readily accessible to everyone. Based on both survey and student activity data, Kizilcec and Halawa found that “the primary obstacle for most [MOOC] learners was finding time for the course” [90]. By conducting post-course surveys, [90] found that 66% of students struggled to keep up with course deadlines and 46% reported that the course required too much time.

Self-regulated learning

Providing income to students in exchange for real-world tasks can serve as a support mechanism in encouraging students to better self-regulate their study and engagement habits. The study of Self-Regulated Learning (SRL) has a rich history in the traditional classroom setting [126, 179], but now the new challenge arises of how to support and enable non-traditional and disadvantaged students to practice effective SRL habits in online/distance learning endeavors. SRL is defined as a student’s proactive engagement with his or her learning process by which various personal organization and management strategies are used in order to control and monitor one’s cognitive and behavioral process towards a learning outcome [157, 178]. Many SRL tactics hinge on effective time management skills [17, 117]. Although, with proper coaching, many students can be taught to find and make time for studies [118, 117], this is simply not plausible for others who do not have enough time in a day to introduce a new challenge–no matter how well they manage their time. These learners are the primary target of our vision. By introducing these opportunities to earn money while completing a course, we hope that they can essentially “buy time.”

For the group of students who complete the paid tasks in order to make “extra” money, the compensation can be viewed as a reward mechanism and
an incentive to prioritize the MOOC over other less important tasks [54, 90]. For the other group, the money earned from the extra tasks is a required means for them to commit time. Whereas reward-seeking students would no longer have a reason to complete the extra tasks if the monetary prize was removed, the other group of students would no longer have the time or the ability.

**Using rewards to motivate learning**

One of the leading critiques of reward programs in traditional education settings is that their prize pool is finite, and once that is exhausted, student motivation will dwindle [165]. In our setup, however, this is not an issue, as online work platforms are consistently replenished with new tasks to recommend to our MOOC learners. This model thus shows the potential for sustainability at scale.

The existing literature on paying or rewarding students with material goods is concerned with young students in traditional classroom settings [46, 58, 67, 165], however the people who stand to benefit the most from the inclusion of freelance projects and tasks into the MOOC environment are predominantly non-traditional students.

[68] approaches the dilemma of incentivising student performance with money through an economic lens. In order to test how financial incentives impact student performance in historically-disadvantaged and under-performing school districts in the United States, this study compared the effectiveness of input-driven versus output-driven reward systems. It was found that incentives based on student input, such as completing assignments or reading books, are more effective than those based on output, such as test scores and grades [46, 58, 84]. In line with the concept of instructional scaffolding, this finding suggests that incentivising and rewarding intermediate tasks along the path to a larger learning goal or objective is more effective than rewarding only the goal itself. Likewise, one of these intermediate tasks especially challenging to open learning is that of allocating and committing time, and we hope the potential to get paid for this time will support learners in doing so.

**Incentives for underprivileged learners**

We also see the introduction of opportunities for learners to contribute to online work market places while taking a MOOC as a potential manner by which we can mitigate belonging uncertainty for under-privileged learners [87, 160]. This is characterized by stigmatized or minority group members feeling uncertain and discouraged by their social bonds in a given environment [160].
If a student sees his or her participation in the course with an immediately clear and relevant purpose—learning the necessary skills to complete this real-world task—then it should thus mitigate any uncertainty or doubt about the students belonging. Walton and Cohen found that interventions designed to reduce/remove feelings of belonging uncertainty can have great effects on students’ subjective experiences in academic settings which can therefore boost academic performance. Learners of low socio-economic status are not the only ones who stand to benefit from this. Other major demographics, such as women (particularly in STEM courses), are currently outnumbered, and often outperformed [74], by their male student counterparts [44, 132, 90].

**Using extra credit to motivate learning**

Many studies have examined the effect that offering extra credit assignments to students can have on overall class performance. [25] found that extra credit assignments can be used to motivate students to read journal articles; [20] found extra credit, in the form of an in-class token economy, to increase course participation; [164] saw increases in course attendance stemming form the offering of extra credit assignments; and [116] found that extra credit assignments can facilitate mastery of course material and strongly predict final exam performance.

Similarly, in a study that specifically targeted students on the verge of failing a college course, researchers found that an intervention in the form of a skills-based extra credit assignment increased these students’ final exam grades, increased and diversified their engagement, and decreased their dropout/incompletion rate [84].

In December 2015, edX, one of the most popular MOOC platforms announced a new policy which rescinds the free honor code course completion certificates previously made available to any student who earned a passing grade in the course. Instead, according to the announcement on the edX blog [4], 'all of edX’s high-quality educational content, assessments and forums will continue to be offered for free, but those learners who want to earn a certificate upon successful completion of the course will pay a modest fee for a verified certificate.' While both edX and its partner institutions will offer various levels of financial aid to students who apply, the design introduced in this work has the potential to reduce the burden of supporting students. Simply by completing one task from an online marketplace (of high enough value), a student can offset the cost of the verified course certificate.

To the best of our knowledge, this effort to pay students in an open learning environment in order to encourage and enable student engagement
is the first of its kind. Research findings in this area promise to help narrow
the established achievement gap we currently observe among MOOC learners.

5.3 EX101x

To investigate our research questions, we inserted bonus exercises, drawn
from paid tasks posted on Upwork, into the MOOC Data Analysis: Take It
to the MAX(), or in short: EX101x. EX101x is a MOOC offered on the edX
platform; its first edition (the one we deployed this study in) ran between
March 31, 2015 and June 18, 2015. The core objective of EX101x is to
learn to conduct data analysis using spreadsheets. Throughout the first six
course weeks, the following set of skills are taught (using Excel as specific
spreadsheet instance): string manipulation and conditional statements (Week
1), lookup and search functions (Week 2), pivot tables (Week 3), named
ranges (Week 4), array formulas (Week 5) and testing in spreadsheets (Week
6). Week 7 is dedicated to the programming language Python and its use
within spreadsheets, while the final week (Week 8) introduces the graph
database Neo4j.

As is common in MOOCs today, learners were invited to participate in a
pre-course and a post-course survey containing questions on the motivation
of the learners, the perceived quality of the course, etc. In September 2015
we approached a selected subset of all learners for an additional post-course
survey.

The course was set up as an xMOOC: lecture videos were distributed
throughout the 8 teaching weeks. Apart from lectures, each week exercises
were distributed in the form of multiple choice and numerical input questions.
Each of the 136 questions was worth 1 point and could be attempted twice.
Answers were due 3 weeks after the release of the respective assignment.
To pass the course, ≥ 60% of the questions had to be answered correctly.
Each week, alongside the usual assignments, we posted one additional bonus
exercise.

Overall, 33,515 users registered for the course. Less than half of all learn-
ers (45%) engaged with the course, watching at least one lecture video. The
completion rate was 6.53% in line with similar MOOC offerings. Over
65% of the learners were male and more than 76% had at least a Bachelor
degree.
5.4 Approach

The design of our experiments was guided by our research questions. As we aim to determine whether learners can solve real-world tasks that are related to the course material with high accuracy and high quality (RQ 4.1), for the six weeks of EX101x that cover data analysis topics in spreadsheets, we manually selected appropriate paid tasks from the Upwork platform — one task per course week. No bonus exercises were posted in weeks 6 and 8 due to the topics covered that week: testing in spreadsheets and the graph database Neo4J. We chose Upwork (which at that time was still called oDesk) as it is one of the largest online work platforms in the English speaking world (cf. Table 5.4); for each course week, we chose an Upwork task that was strongly related to that week’s course content by extensively scanning the currently active Upwork tasks worth up to $50. We chose this price limit to provide tasks that can be solved in a reasonable amount of time. We kept the task description intact, and added a short introduction to provide the necessary context to our learners (i.e. a clear disclaimer that this is a real-world task). A concrete example of a bonus exercise derived in this manner is shown in Figure 5.2; it was posted in week 4 of EX101x.

To answer RQ 4.2 and RQ 4.3, we explored the suitability of Upwork as a source of paid tasks along several dimensions including the covered topics, the task longevity, and the financial gain. In order to investigate RQ 4.1 and RQ 4.4 we require exact definitions of a number of metrics (i.e. accuracy, coverage, quality and engagement). In the following section, we describe them in detail.

5.4.1 Measurements

Accuracy. For each bonus exercise, we developed a gold standard solution in collaboration with the course instructor and verified whether the submitted learner solutions matched the gold standard solution, thus measuring their accuracy. We considered a submitted spreadsheet a match to our gold standard if it contained the required solution columns with the correct cell content; additional columns were ignored; slight deviations from the gold standard (e.g. an empty string or “N/A” instead of an empty cell in the gold standard) were allowed. We iteratively refined our automated grading script by randomly sampling 20 submission in each iteration (and manually verifying the correctness of the grading script) until all samples were classified correctly.
5.4. Approach

Have you ever sold anything on Amazon.com? For this real-world task (again derived from an actual oDesk task), we put you in the shoes of an Amazon seller who is selling accessories for pets. The seller himself buys these accessories from a supplier. The seller currently has a five star feedback rating on Amazon. To keep it this way, only items that the seller can immediately ship should appear in the seller’s Amazon storefront (i.e., those items that the supplier has in stock).

The seller has this Excel sheet which stores the ID of all products to be posted on his Amazon.com storefront and the number of units available, as illustrated in the example below.

<table>
<thead>
<tr>
<th>ID</th>
<th>Stock</th>
</tr>
</thead>
<tbody>
<tr>
<td>08357E0A-DF5C-392E-F7CC-A27E3AB7680DF</td>
<td>?</td>
</tr>
<tr>
<td>08357E0A-DF5C-392E-F7CC-DD44553D03DEF</td>
<td>?</td>
</tr>
</tbody>
</table>

File 1: Excel sheet of the seller

It is your job to update the Stock column based on the information the seller receives from the supplier.

Every day, the seller receives an Excel sheet from his supplier, which contains the supplier’s inventory. An example is provided below. Note that the supplier’s column Product corresponds to the seller’s column ID.

<table>
<thead>
<tr>
<th>Product</th>
<th>Inventory</th>
</tr>
</thead>
<tbody>
<tr>
<td>08357E0A-DF5C-392E-F7CC-A27E3AB7680DF</td>
<td>?</td>
</tr>
<tr>
<td>08357E0A-DF5C-392E-F7CC-DD44553D03DEF</td>
<td>188</td>
</tr>
</tbody>
</table>

File 2: Excel sheet of the supplier

To keep his customers satisfied, the seller uses the following two rules to set the Stock column:

- If the supplier’s inventory of a product is less than 30, Stock should be set to 0;
- If the supplier’s inventory of a product is more than or equal to 30, Stock should be set to 20.

Applying these two rules to our example files above, yields the following result:

<table>
<thead>
<tr>
<th>ID</th>
<th>Stock</th>
</tr>
</thead>
<tbody>
<tr>
<td>08357E0A-DF5C-392E-F7CC-A27E3AB7680DF</td>
<td>0</td>
</tr>
<tr>
<td>08357E0A-DF5C-392E-F7CC-DD44553D03DEF</td>
<td>20</td>
</tr>
</tbody>
</table>

File 3: Result sheet of the seller

Please send your solutions to ...

Figure 5.2: Bonus exercise posted in week 4 of EX101x. The original task was posted with a price of $35 to Upwork (note that at the time of posting this exercise, Upwork was still called oDesk).

Coverage. Besides accuracy, we also measured the coverage of learner solutions. We operationalize coverage as the percentage of cells that the learner
solution shares with the gold standard. As for accuracy, we ignored additional columns and allowed minor deviations in the cells such as additional white spaces or minimal numeric differences to account for floating point inaccuracy on different computers. Coverage can be seen as an indicator of how close the solution is to the gold standard solution.

**Quality.** To investigate the quality of the submissions, we turned to the concept of code smells [153], an established measure of quality in the field of Software Engineering: code smells are specific to particular programming languages; spreadsheets code smells include standard errors (e.g., #N/A!, #NAME?), high conditional complexity (e.g., involving too many nested IF operations), hidden rows/columns/worksheets, etc. We adopted the code smells for spreadsheets proposed in [71] and rank the solutions by the number of smells they exhibit - the fewer smells a solution has, the higher its quality.

**Engagement.** Finally, based on our experimental setup, we are also able to investigate the effect of real-world tasks on student engagement (RQ 4.4). We hypothesize that learners who view the bonus exercises and realize that those are real-world tasks that could earn them money, will become more engaged with the course material than learners who did not view the bonus material. To this end, we only consider the subset of active learners \( L_{\text{noBonus}} \) that did not submit any solutions to the bonus exercises.

We group learners together that are similarly engaged in the course up to the point of either viewing a bonus exercise or not. If our hypothesis holds, then after that point in time, those learners that viewed the bonus exercise should, on average, exhibit higher engagement than those that did not.

We operationalize this experiment as follows: we measure a learner’s engagement through his or her amount of video watching. In week 1, we partition the learners in \( L_{\text{noBonus}} \) in two groups: we sort the learners in video watching time order and then split them in two equally sized groups - the lower half is the low engagement, and the upper half is the high engagement group. We then compute for each learner the amount of video watching in all following weeks and determine for the low and high engagement groups separately whether there is a statistically significant difference between those learners that did view and those that did not view the bonus exercise. In week 2, we repeat this analysis by taking as starting point only the subset of learners in \( L_{\text{noBonus}} \) that viewed the bonus exercise in 1. We repeat those steps until week 7 (in each week resorting the remaining learners into the low and high engagement groups). While we expect significant differences based on bonus exercise viewing in the early weeks of the course, we should
not observe significant differences towards the end of the course — in week $n$ we only include learners that up to that point in time have viewed all $n - 1$ bonus exercises. At some point, bonus exercises should not provide additional engagement anymore.

5.5 Results

Before we discuss our results for each of the four research questions in turn, we provide a first global view of our learner population in EX101x.

<table>
<thead>
<tr>
<th></th>
<th>All Engaged Learners</th>
<th>BE Learners</th>
<th>Non-BE Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Learners</td>
<td>15,074</td>
<td>2,020</td>
<td>13,054</td>
</tr>
<tr>
<td>Completion rate</td>
<td>14.02%</td>
<td>44.11%</td>
<td>9.36%</td>
</tr>
<tr>
<td>Avg. time watching</td>
<td>58.78</td>
<td>133.48</td>
<td>47.21</td>
</tr>
<tr>
<td>video material (in min.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%Learners who tried at least one question</td>
<td>59.89%</td>
<td>98.56%</td>
<td>53.91%</td>
</tr>
<tr>
<td>Avg. #questions learners attempted to solve</td>
<td>24.06</td>
<td>67.41</td>
<td>17.36</td>
</tr>
<tr>
<td>Avg. #questions answered correctly</td>
<td>19.56</td>
<td>55.60</td>
<td>13.98</td>
</tr>
<tr>
<td>Avg. accuracy of learners’ answers</td>
<td>53.40%</td>
<td>90.09%</td>
<td>47.73%</td>
</tr>
<tr>
<td>#Forum posts</td>
<td>10,106</td>
<td>4,341</td>
<td>5,765</td>
</tr>
<tr>
<td>%Learners who posted at least once</td>
<td>16.20%</td>
<td>43.61%</td>
<td>11.96%</td>
</tr>
<tr>
<td>Avg. #posts per learner</td>
<td>0.67</td>
<td>2.15</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 5.1: Basic characteristics across all learners and their partitioning into those who attempted to solve at least one Bonus Exercise (BE) and those who did not (Non-BE). Where suitable, significance tests between the BE/Non-BE groups were performed according to Mann-Whitney. All performed tests exhibited significant differences - indicated with ‡ (significant difference with $p < 0.001$).

We classified our set of engaged learners, i.e., those who watched at least one video² (a definition also employed for instance in [65]), according to two

---

²We note, that we also evaluated two alternative definitions of engagement: (1) learners that watched at least 15 minutes of video material (i.e. at least two videos), and (2) learners that submitted at least five quiz questions. While the absolute values reported in Tables 1
dimensions: (i) whether learners attempted to solve at least one bonus exercise (BE) or not (Non-BE) and (ii) the number of bonus exercises learners attempted to solve. In the latter case, we consider only the BE learners. We mark learners as dedicated bonus exercise solvers (DBE) if they attempted to solve more than two bonus exercises, the remaining learners are non-dedicated (Non-DBE). The basic statistics of both learner cohorts are presented in Tables 5.1 and 5.2. It is evident that learners who solved at least one bonus exercise are more engaged than learners who did not - across all important characteristics (average time spent watching videos, average number of questions answered, accuracy of answers) the BE learners perform significantly better than the Non-BE learners. Among the cohort of BE learners, this trend continues with the dedicated learner group being significantly more engaged and successful than the non-dedicated learner group.

We note that these results are not surprising — they are dictated by common sense and our manner of classifying learners. Importantly, we do not claim a causal relationship between bonus exercise presence and learner engagement based on these results (in Section 5.5.3 we explore the relationship between engagement and bonus exercises in greater detail).

As our goal is to improve the ability of learners from the developing world to engage and successfully complete the course, we also investigate to what extent they are already capable of doing so now. For each country, we computed the percentage of learners that completed the course (based on all registered learners). Shown in Figure 5.3 is the completion rate of EX101x across countries, split into developed countries according to the OECD (in blue) and developing countries (in red). We observe, that in general, the completion rate of learners from developed countries is higher than those of developing countries (with the exception of Russia and Malaysia). This confirms one of our assumptions that learners from developing countries are facing issues that learners in developed countries do not face. This result is in line with previous findings in [36].

5.5.1 RQ 4.1: Can learners solve real-world tasks well?

Across all weeks, we received a total of 3,812 bonus exercise solutions from 2,418 learners. Since the edX platform has very limited solution uploading capabilities, we asked learners to email us their solutions and then matched the email addresses of the learners to their edX accounts. 352 of the learners & 2 change depending on the definition employed, we did observe the same trends and the same significant differences for all three engagement definitions and thus only report one.
5.5. Results

<table>
<thead>
<tr>
<th></th>
<th>DBE Learners</th>
<th>Non-DBE Learners</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Enrolled learners</td>
<td>314</td>
<td>1,706</td>
</tr>
<tr>
<td>Completion rate</td>
<td>86.31%</td>
<td>36.34%</td>
</tr>
<tr>
<td>Avg. time watching</td>
<td>189.45</td>
<td>123.18</td>
</tr>
<tr>
<td>video material (in min.)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Learners who tried</td>
<td>100.00%</td>
<td>98.30%</td>
</tr>
<tr>
<td>at least one question</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. # questions learners attempted to solve</td>
<td>110.52</td>
<td>59.47</td>
</tr>
<tr>
<td>Avg. # questions answered correctly</td>
<td>93.99</td>
<td>48.53</td>
</tr>
<tr>
<td>Avg. accuracy of learners’ answers</td>
<td>94.83%</td>
<td>89.22%</td>
</tr>
<tr>
<td>#Forum posts</td>
<td>1,626</td>
<td>2,715</td>
</tr>
<tr>
<td>% Learners who posted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>at least once</td>
<td>59.87%</td>
<td>40.62%</td>
</tr>
<tr>
<td>Avg. # posts per learners</td>
<td>5.18</td>
<td>1.59</td>
</tr>
</tbody>
</table>

Table 5.2: Basic characteristics of BE learners partitioned into dedicated BE learners (DBE) solving 3+ bonus exercises and non-dedicated BE learners. Where suitable, significance tests between the DBE/Non-DBE groups were performed according to Mann-Whitney. All performed tests exhibited significant differences - indicated with ‡ (significant difference with $p < 0.001$).

could not be matched to an edX account (i.e. these learners used a different email when signing up for edX) and had to be excluded from the subsequent analyses of edX log traces (they are included though in all results analyzing the accuracy/quality of the solutions).

Table 5.3 lists the main results of our accuracy and quality analyses. Between 1% (in week 7) and 15% (in week 1) of active learners participated in the bonus tasks each week. The percentage of accurate solutions varies widely between tasks and is not correlated with the amount of pay for a task. In fact, the two tasks with the lowest pay ($20 in weeks 3 & 5) resulted in the lowest percentage of accurate solutions (11% and 17% respectively). The low accuracy for the seemingly simple (as cheaply priced task) is intriguing. We sampled 50 of the incorrect solutions and found most of them to miss a required final step in the task. Both tasks require students to carefully read and understand the assignment to be successful. In week 3, learners needed to implement an equation containing an absolute value. As the equation...
text is fairly long, students tended to miss this vital piece of information; 78% of all wrong answers that week show this misconception. In week 5, the solutions had a similar issue, often missing a final re-ranking step of the result columns as required in the task description.

An alternative view of submission accuracy is presented through the average coverage of all submissions, that is the fraction of gold standard result cells, that were also present in the submissions. Coverage is 1.0 for the correct submissions, but usually lower for incorrect ones (note that it is possible for an incorrect solution to reach a coverage of 1.0 if it contains all gold standard result cells as well as additional result cells - this happens rarely though). In Table 5.3 we observe that the coverage across all submitted solutions is rather high (with the exception of week 3), thus even solutions that are not correct are at least sensible.

Having considered accuracy and coverage, we now turn to the quality of the solutions. Among the correct solutions, a large fraction (between 38% and 96%) are of high quality, that is they exhibit zero code smells as shown in Table 5.3. Again, we do not observe a correlation between the price of a task and the quality of the solutions. The quality of the accurate and inaccurate solutions (as measured in code smells) is comparable. Across all weeks and submitted solutions, the median number of code smells is less than 10, indicating that most learners were able to code high-quality solutions.
5.5. Results

The vast majority of solutions across all weeks have less than 50 reported code smells.

Overall, we can positively answer RQ 4.1: it is indeed possible for MOOC learners to provide correct and high-quality solutions to selected real-world tasks from an online work platform.

<table>
<thead>
<tr>
<th>Week</th>
<th># Active learners</th>
<th># Bonus learners (% from active)</th>
<th>Task payment</th>
<th># Accurate solutions (% of active)</th>
<th># High quality solutions (% of accurate)</th>
<th>Coverage (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13,719</td>
<td>2,145 (15.64%)</td>
<td>$25</td>
<td>1,731 (80.70%)</td>
<td>1,230 (71.06%)</td>
<td>0.88 (0.32)</td>
</tr>
<tr>
<td>2</td>
<td>8,228</td>
<td>594 (7.22%)</td>
<td>$50</td>
<td>227 (38.22%)</td>
<td>87 (38.33%)</td>
<td>0.91 (0.27)</td>
</tr>
<tr>
<td>3</td>
<td>5,825</td>
<td>390 (6.70%)</td>
<td>$20</td>
<td>44 (11.28%)</td>
<td>28 (63.64%)</td>
<td>0.54 (0.32)</td>
</tr>
<tr>
<td>4</td>
<td>4,270</td>
<td>414 (9.70%)</td>
<td>$35</td>
<td>354 (85.51%)</td>
<td>296 (83.62%)</td>
<td>0.95 (0.22)</td>
</tr>
<tr>
<td>5</td>
<td>3,709</td>
<td>231 (6.23%)</td>
<td>$20</td>
<td>39 (16.88%)</td>
<td>16 (41.03%)</td>
<td>0.69 (0.24)</td>
</tr>
<tr>
<td>7</td>
<td>3,059</td>
<td>38 (1.24%)</td>
<td>$35</td>
<td>26 (68.42%)</td>
<td>25 (96.15%)</td>
<td>0.73 (0.68)</td>
</tr>
</tbody>
</table>

Table 5.3: Learners' performance on real-world tasks. The second column shows the number of active learners. The third column shows the number of students taking the bonus exercise. The fourth column shows the task payment offered at UpWork. Accurate submissions are those matching our gold standard (with the additional requirement of the correct order for tasks 3 and 5). High-quality submissions are those correct submissions without code smells. The coverage column reports the average (and standard deviation) fraction of cells covered by all of a week’s submissions.

5.5.2 RQ 4.2 & RQ 4.3: An exploratory analysis of UpWork

We first note that Upwork is only one of multiple large online work platforms in the English speaking world as shown in Table 5.4. Together those companies facilitated more than 2.5 billion dollars in worker payments. Important for us, some of these platforms (including Upwork) provide API access to their content, thus enabling a recommender system as we envision.

For our analysis, we took a snapshot of all available tasks on Upwork on September 15, 2015 leading to a total of 56,308 open tasks. Each task is assigned to one or more topical categories, e.g. Translation or IT & Networking. Additionally, tasks can be tagged with particular required skills such as excel or python. Tasks either pay per hour or have a fixed budget. We focus on the latter, as the budget is a direct indicator for the amount of work required. A task pays on average $726 (SD: $3,417) and stays 27 days on the platform (SD: 34 days) before being solved or canceled. Among all tasks, we found 574 spreadsheet tasks (potentially relevant for EX101x) in the budget range from $1 - $50. A task in this (budget) subset stayed 25 days on the platform on average (SD: 40 days).
To estimate the proportion of tasks that may be suitable recommendations for EX101x learners, we analysed a random sample of 80 tasks of the budget set. An expert classified these tasks into three categories:

1. *lecturable* are tasks that would make them suitable as course material for a specific lecture (e.g. a task that requires knowledge of a spreadsheet’s VLookUp function);

2. *relevant* are tasks that fit the topic yet do not fit into a specific lecture (e.g. a task that requires the use of spreadsheets but otherwise does not rely on knowledge taught in the course);

3. *unrelated* are all other tasks that do not fit in the courseware in general.

Among the 80 tasks we found 34 unrelated tasks, 39 relevant tasks and 7 lecturable tasks. Based on these numbers and the average time a task stays online we can estimate how many tasks are added every day to Upwork that fit our criteria (i.e. have a price between $1 and $50 and require spreadsheet knowledge): 10 unrelated tasks, 11 relevant tasks, and 2 lecturable tasks. These numbers indicate that there are not yet enough budget tasks available to provide individual MOOC learners with weekly opportunities to earn money whilst learning — at least for the EX101x MOOC.

<table>
<thead>
<tr>
<th>Company</th>
<th>Paid worker fees</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upwork</td>
<td>$ 1,000 M</td>
<td>yes</td>
</tr>
<tr>
<td>witmart</td>
<td>$ 1,000 M</td>
<td>no</td>
</tr>
<tr>
<td>freelance</td>
<td>$ 462 M</td>
<td>no</td>
</tr>
<tr>
<td>Guru</td>
<td>$ 200 M</td>
<td>yes</td>
</tr>
<tr>
<td>Envato</td>
<td>$ 200 M</td>
<td>yes</td>
</tr>
<tr>
<td>Topcoder</td>
<td>$ 72 M</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table 5.4: Paid total worker fees by company in Million US Dollar. These numbers are self reported by the companies and are not given for a specific year.

One limiting factor in our design is the budget limit we set ourselves ($50). The majority of tasks have a higher budget as shown in Figure 5.2 and future experiments will investigate the question up to which budget level learners are able to solve tasks in a reasonable amount of time, with high accuracy and high quality.

Tasks that have a higher budget (on the topic of spreadsheets) are usually more intricate and instead of solving one specific problem in a spreadsheet
Figure 5.4: From the 56,308 Upwork tasks available on 15/09/2015 a total of 8,153 have a fixed budget (the remaining tasks are paid by the hour). Budgeted tasks are binned according to the budget they have.

(as less pricey tasks, cf. Figure 5.2) they often require the development of a complete solution as exemplified in the three task examples priced between $100 and $500 at Upwork:

**$500** “We are commercial real estate brokers and are looking for an expert in Microsoft Excel to create an interactive Excel worksheet(s) for rental comparison purposes.”

**$250** “I need to have financial calculations for a customer equity/lifetime value model integrated into an excel workbook. (…)”

**$100** “I currently plot support and resistance zones manually on a chart like the attached image. (…) I need to calculate these support and resistance levels within MS Excel programmatically or using some sort of algorithm. (…)”

In contrast to the budget, the longevity of tasks on Upwork is beneficial for our vision. Figure 5.5 shows that many tasks remain available for at least 20 days, which is beneficial in the MOOC setting where assignments also commonly have a grace period of 2-3 weeks.

Recall, that additionally to a general category each task is tagged with a set of required skills. Table 5.5 shows Excel (the comon tag for spreadsheet
Figure 5.5: The 56,308 Upwork tasks available on 15/09/2015 are binned according to the number of days they have been “online” (i.e. the task is open).

tasks) to be a relatively popular task. More general skills such as proficiency in HTML and CSS occur more often than specific skills such as proficiency in R. Overall, programming tasks only make up a small percentage of all available tasks, as shown in Table 5.1. Indeed, the breadth of tasks offered on Upwork indicates the tremendous potential of online work platforms for suggesting paid tasks to learners across a range of MOOCs.

To conclude, we observe that, indeed, the knowledge gained during EX101x can be used to solve paid tasks (RQ 4.2), though the number of tasks posted per day that fit our criteria is rather low: we estimate that, on average, 13 tasks a day are posted in the $1-$50 category, requiring spreadsheet knowledge fitting the course topic of EX101x.

This result also provides an answer to RQ 4.3 in the context of EX101x: as per day, on average, only 13 MOOC students stand to benefit from these paid tasks (i.e., can earn money from them), there are not sufficient tasks available to sustain a standard MOOC population of learners throughout an entire run of EX101x — at least at the current rate of online work tasks being posted to Upwork.
5.5. Results

<table>
<thead>
<tr>
<th>#Tasks</th>
<th>Skill Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,443</td>
<td>HTML5 &amp; HTML</td>
</tr>
<tr>
<td>5,034</td>
<td>PHP</td>
</tr>
<tr>
<td>3,928</td>
<td>Javascript</td>
</tr>
<tr>
<td>2,731</td>
<td>Excel</td>
</tr>
<tr>
<td>616</td>
<td>Python</td>
</tr>
<tr>
<td>559</td>
<td>Ruby &amp; Ruby-on-Rails</td>
</tr>
<tr>
<td>537</td>
<td>Objective-c</td>
</tr>
<tr>
<td>450</td>
<td>Java</td>
</tr>
<tr>
<td>26</td>
<td>Perl</td>
</tr>
<tr>
<td>34</td>
<td>R</td>
</tr>
</tbody>
</table>

Table 5.5: Overview of programming tasks among our crawl of 56,308 Upwork tasks on 15/09/2015.

<table>
<thead>
<tr>
<th>Category</th>
<th>#Tasks</th>
<th>Days Online (SD)</th>
<th>Payment (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer Service</td>
<td>986</td>
<td>74.76 (83.24)</td>
<td>$1,817 ($6,692)</td>
</tr>
<tr>
<td>Engineering &amp; Architecture</td>
<td>1,432</td>
<td>53.70 (61.50)</td>
<td>$1,699 ($6,640)</td>
</tr>
<tr>
<td>Translation</td>
<td>2,109</td>
<td>53.02 (74.64)</td>
<td>$1,156 ($3,710)</td>
</tr>
<tr>
<td>Admin Support</td>
<td>5,961</td>
<td>50.33 (89.14)</td>
<td>$ 982 ($4,855)</td>
</tr>
<tr>
<td>Accounting &amp; Consulting</td>
<td>1,095</td>
<td>49.37 (77.77)</td>
<td>$ 997 ($4,642)</td>
</tr>
<tr>
<td>IT &amp; Networking</td>
<td>2,182</td>
<td>39.60 (52.38)</td>
<td>$ 854 ($4,356)</td>
</tr>
<tr>
<td>Data Science &amp; Analytics</td>
<td>1,156</td>
<td>37.29 (45.94)</td>
<td>$ 777 ($3,308)</td>
</tr>
<tr>
<td>Writing</td>
<td>8,448</td>
<td>32.31 (58.54)</td>
<td>$ 418 ($ 832)</td>
</tr>
<tr>
<td>Legal</td>
<td>333</td>
<td>27.97 (33.67)</td>
<td>$ 377 ($2,055)</td>
</tr>
<tr>
<td>Web, Mobile &amp; Software Dev</td>
<td>16,328</td>
<td>25.39 (46.02)</td>
<td>$ 376 ($2,028)</td>
</tr>
<tr>
<td>Design &amp; Creative</td>
<td>9,667</td>
<td>24.60 (45.70)</td>
<td>$ 274 ($ 710)</td>
</tr>
<tr>
<td>Sales &amp; Marketing</td>
<td>6,724</td>
<td>21.54 (34.31)</td>
<td>$ 263 ($2,124)</td>
</tr>
</tbody>
</table>

Table 5.6: The 56,308 Upwork tasks available on 15/09/2015 are partitioned according to their category. Shown are the number of tasks per category, the average number of days online and the average task payment (for the subset of 8,153 tasks with a fixed budget).

5.5.3 RQ 4.4: Learner engagement

We hypothesize that our bonus exercises, in particular the realization that those are real-world tasks with which money could be earned, are beneficial for learner engagement.

In Figure 6.6 we present the results of our experiment, comparing the amount of video watching between learners who did view and did not view
the bonus exercises (computed separately for low and high engagement learners). Let’s consider week 1: in the low engagement group, the learners that did not view the bonus exercise spent on average 0.08 hours (5 minutes) in subsequent weeks on video watching, while the learners that did view the bonus exercise spent 1.3 hours in subsequent weeks on videos. This difference is statistically significant ($p < 0.001$, Mann-Whitney test). Similarly, in the high engagement group, learners that did not view the bonus exercise continued to spend 0.4 hours (24 minutes) on video watching, while learners that did view the bonus exercise spent 1.7 hours on the course. Across both engagement groups, the low amount of overall time spent in watching videos can be explained by the fact that over time, more and more learners drop out of a course. In week two, we only consider the subset of learners that viewed the bonus exercise in week 1, and again we observe significant differences in engagement between those that viewed the second bonus exercise and those that did not. As the weeks go on, the difference in video watching time between learners viewing and not viewing the bonus exercise of the week tends to decrease—also evident in the fact that in weeks 5 and 7, we find no significant differences in engagement for the high engagement learners. We consider these results as a first confirmation of RQ 4.4: our bonus exercises (real-world tasks) are likely to have a positive effect on engagement. We realize that this experiment can only be considered as first evidence: we observed that similarly engaged learners diverge in their behavior after having (not) viewed our real-world bonus tasks. We assume that this divergent behavior is caused by the action of (not) viewing the task, but this assumption cannot be directly verified. We attempt to verify it (among others) through a post-course survey, outlined next.

5.5.4 Post-course survey

We sent a follow-up survey with 11 questions (about success & engagement in EX101x, financial incentives in MOOC learning and the bonus tasks in EX101x) to a subset of learners who expressed their willingness to be contacted after the course had completed. An overview of all questions can be found in Table 5.7.

We partitioned the set of contacted learners into four groups according to their origin (developed vs. developing country) and their engagement with the bonus exercises (submitted vs. not submitted):

- from developed nations & submitted at least one bonus exercise (126 learners contacted, 26 replied);
## 5.5. Results

### Success & engagement

<table>
<thead>
<tr>
<th>Question</th>
<th>Completed successfully</th>
<th>Stopped midway</th>
<th>Stopped in weeks 1 or 2</th>
<th>Registered, but nothing else</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>How engaged were you in EX101x?</td>
<td>74 (87%)</td>
<td>6 (7%)</td>
<td>8 (10%)</td>
<td>1 (1%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Submitted 3+ exercises</th>
<th>Submitted 1-2 exercises</th>
<th>Attempted 1+ exercises, not submitted</th>
<th>Looked at 1+ exercises, not attempted</th>
<th>Know about exercises, did not look at any</th>
<th>No knowledge of exercises</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>84 (84%)</td>
<td>41 (41%)</td>
<td>11 (11%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Not applicable</th>
<th>No</th>
<th>Yes, a minor factor</th>
<th>Yes, a major factor</th>
<th>-</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>74 (74%)</td>
<td>26 (26%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Financial incentives in general

<table>
<thead>
<tr>
<th>Question</th>
<th>No financial incentive required</th>
<th>$0-$9</th>
<th>$10-$29</th>
<th>$30-$49</th>
<th>$50-$99</th>
<th>$100+</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80 (80%)</td>
<td>48 (48%)</td>
<td>32 (32%)</td>
<td>11 (11%)</td>
<td>5 (5%)</td>
<td>5 (5%)</td>
</tr>
</tbody>
</table>

### Bonus exercises in EX101x

<table>
<thead>
<tr>
<th>Question</th>
<th>Not applicable</th>
<th>1 (Essential)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 (Extra)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80 (80%)</td>
<td>48 (48%)</td>
<td>66 (66%)</td>
<td>8 (8%)</td>
<td>9 (9%)</td>
<td>8 (8%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Open-answer form</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Open-answer form</th>
</tr>
</thead>
</table>

### Table 5.7: Overview of the 11 questions in our post-course survey. For presentation purposes, some questions and answers appear slightly condensed. For all closed-form questions, we provide the distribution of answers (in %) across the four learner partitions in the form A | B | C | D%: (A) from developed nations + at least one bonus exercise submitted, (B) from developing nations + at least one bonus exercise submitted, (C) from developed nations + no bonus exercise submitted, and, (D) from developing nations + no bonus exercise submitted.
Chapter 5. Enabling MOOC Learners to Solve Real-world Paid Tasks

Figure 5.6: The average amount of time (in hours) that learners spent in watching video after viewing (but not submitting) the bonus exercises. The numbers of learners within each group are given in brackets. Results marked with * ($p < 0.001$) are significantly different (Viewed vs. Not viewed) according to the Mann-Whitney U-test.

- from developing nations & submitted at least one bonus exercise (114 learners contacted, 29 replied);
- from developed nations & did not submit a bonus exercise (357 learners contacted, 34 replied);
- from developing nations & did not submit a bonus exercise (271 learners contacted, 22 replied);

Besides the questions and answer options, in Table 5.7 we also report the distribution of given answers for all closed-form questions and each learner partition. We note that a small number of learners who we classified as not
5.5. **Results**

having submitted a bonus solution self-reported having done so. The converse is also true: a small number of learners that we have received bonus exercise submissions from reported not having submitted any. These self-reporting errors could be explained by the amount of time (12 weeks) passed between the end of EX101x and the release of the survey. Overall though, the vast majority of learners were remembering their (lack of) submissions for our bonus exercises correctly.

Students from developing nations who did not attempt any of the bonus exercises report that if they could earn somewhere between $10 and $100 per week through such online work platform tasks, they would commit up to six more hours to the course per week. In this same group, 45% of respondents attempted one or more bonus exercises but did not submit it to the course instructor. In contrast, of the survey respondents from developed nations who did not submit a bonus exercise to the instructor, only 20% reported having attempted to solve any. This difference suggests that learners from developing nations are more motivated and eager to engage with course material, but there seems to be a barrier stopping them from fully engaging as much as they would like. Providing an opportunity for them to gain income in the process could be a key factor in enabling them to fully commit to a MOOC.

In question 9 we asked students how difficult they found the bonus exercises to be on a five-point Likert scale—“1” being too easy and “5” being too difficult. Of the entire group of learners (across all partitions) that responded, the average score was 3.48. As bonus exercises, they are expected to be slightly more difficult than the rest of the course material, and the students seem to generally view them as such—slightly more difficult, yet accessible. This sentiment is also echoed in the students’ comments in the survey when asked why they chose to engage with the bonus exercises in the first place; the three most common words to appear in the responses, in order, are “challenge,” “real,” and “test.” To synthesize, students generally see these activities as an added challenge in which they test their ability to apply what they learned in the course to a real-world problem.

Also interesting is that learners from developing countries perceived the bonus exercises as being more difficult than learners in developed countries (Mann-Whitney U-test with $U = 781$, $Z = -2.13$ and $p < 0.05$). This discrepancy underlines the importance for learners in developing countries to be able to commit the necessary time for these types of tasks, as a higher perceived difficulty would require more time from the learner to understand and/or master the content.
Finally, we also explored the effect of the bonus exercises on learners’ motivation to engage with the course (survey question 8). These responses, also on a five-point Likert scale, ranged from “Not at all” (1) to “Very much” (5). A difference emerged in the way learners from different backgrounds are affected by the presence of the bonus exercises. Learners from developing nations report that bonus exercises increased their motivation to engage with the course significantly more than learners from developing countries (Mann-Whitney U-test with $U = 617.5$, $Z = 2.61$ and $p < 0.05$).

### 5.6 Freelance Recommender System Design

Based on our analyses presented in the previous sections, we have to take the following two requirements into account when designing our freelance task recommender:

- The recommender should support multiple task platforms, as we have found Upwork (at this point in time) to only offer a very limited number of tasks in our specified price range and on our specific MOOC’s topic each day.

- Once we recommend learners tasks on Upwork and other similar platforms, we need to continuously track the tasks’ status (are they still available?) as well as the number of times we have recommended them to different learners (to avoid hundreds of learners trying to “bid” for the same task — only one of them can get the job and be paid).

Figure 5.7 shows our designed recommender system, which — for any given MOOC — will automatically retrieve real-world tasks relevant to the topics covered in the MOOC and recommend them to our learners. We briefly discuss the different layers in turn:

- **MOOC.** The MOOC layer serves as the playground for learners to interact with course components and our freelance task recommender system.

- **Data layer.** This layer is responsible for collecting learners’ activity data and gathering real-world tasks from freelance platforms. To be specific, the component MOOC data collector collects data of learners’ interactions with course components (e.g., watching lecture videos, viewing forum posts, submitting quiz answers) and the recommender
Figure 5.7: Overview of the freelance work task recommender system’s design.

system (e.g., viewing recommended freelance tasks, dwell time). On the other side, the component *Freelance task collector* retrieves course-relevant tasks from multiple freelance platforms including *Upwork*, *witmart*, *Guru* and *Envato*. As some of the discovered freelance tasks may not be suitable for our setting of “earning whilst learning” (high budget tasks often require deep knowledge of several fields), the *Task filtering* component filters out unsuitable tasks by applying rule-based strategies (e.g., by setting the maximum budget). In addition, the *Task availability tracker* component regularly checks whether the recommended
freelance tasks are still open & available before generating the recommendations for our learners.

- **Analysis layer.** In this layer, the Learner profiling component analyzes learners’ interaction patterns with the recommender system and how/whether learners’ course engagement can be influenced by freelance task recommendations. The Task relevance estimation component computes the relevance of the discovered tasks with respect to the specific MOOC as well as (potentially) the learner profile.

- **Intervention layer.** At last, the intervention layer makes task recommendations to our learners. The Recommendation diversification component is responsible for presenting a diverse selection of recommendations (to avoid hundreds or thousands of learners competing for the same freelance tasks).

In future work we will implement this design and test its influence in various MOOCs by exploring its effect on MOOC learners.

### 5.7 Conclusion

Can MOOC learners be paid to learn? We set out to provide a first answer to this question in the context of the EX101x MOOC. We found that indeed, work tasks of up to $50 can be solved accurately and in high quality by a considerable percentage of learners that attempt it. We also explored the suitability of the online work platform Upwork in providing tasks to MOOC learners - while there are many budget tasks available (between $1 and $50), those specific to EX101x are rather low in number; at the moment we expect no more than 13 suitable tasks (i.e. specific to taught course material) to be posted per day. Finally, we investigated the matter of engagement: does knowing that real-world tasks may be solved with course knowledge increase learners’ engagement? Our evidence suggests that this is may indeed be the case. We note that while we did observe correlational relationships between learners’ bonus exercise engagement and in-course behavior, the present research cannot yet claim any causality.

Based on the work presented here, we will explore several promising directions (beyond the development and deployment of the presented recommender design) in the future. We will investigate (i) experimental setups that allow us to further investigate the causal relationship between real-world tasks and learner engagement, (ii) the suitability of more complex tasks (i.e.
tasks with a budget greater than $50) for MOOC learners, (iii) the acceptance of the “learners can be earners” paradigm in different populations, and (iv) setups that aid MOOC learners to take the first steps in the paid freelance task world, inspired by [116].
Chapter 6

LearningQ for Educational Question Generation

In this chapter, we focus on the collection of a large-scale and high-quality educational question dataset, as the first step to construct an automatic question generator to ease the burden of instructors in manually creating a suitably large question bank. To this end, we examine the learning material accumulated in two mainstream education-oriented Social Web platforms, i.e., TED-Ed and Khan Academy, and present LearningQ, which consists of a total of 230K document-question pairs, whose questions are of all cognitive levels in the Bloom’s Revised Taxonomy \[104\] and covering various learning subjects. To show the research challenges in generating educational questions, we use LearningQ as a testbed to examine the performance of two state-of-the-art approaches in automatic question generation and investigate possible strategies to select question-worthy sentences from an article. The contributions of this chapter have been published in \[33\].
6.1 Introduction

In educational settings, questions are recognized as one of the most important tools not only for assessment but also for learning [128]. Questions allow learners to apply their knowledge, to test their understanding of concepts and ultimately, to reflect on their state of knowledge. This, in turn, enables learners to better direct their learning effort and improve their learning outcomes.

Previous research has shown that the number of questions learners receive about a knowledge concept is positively correlated with the effectiveness of knowledge retention [8]. It is thus desirable to have large-scale question banks for every taught concept to better support learners.

Designing a suitably large set of high-quality questions is a time-consuming and cognitively demanding task. Instructors need to create questions of varying types (e.g., open-ended, multiple choice, fill-in-the-blank), varying cognitive skill levels (e.g., applying, creating) and varying knowledge dimensions (e.g., factual, conceptual, procedural) that are preferably syntactically different yet semantically similar in order to enable repeated testing of a knowledge concept. To ease instructors’ burden, automatic question generation has been proposed and investigated by both computer scientists and learning scientists to automate the question creation process through computational techniques [110, 136, 137, 69].

Typically, automatic question generation has been tackled in a rule-based manner, where experienced teachers and course instructors are recruited to carefully define a set of rules to transform declarative sentences into interrogative questions [161, 5, 69]. The success of these rule-based methods is heavily dependent on the quality and quantity of the handcrafted rules, which rely on instructors’ linguistic knowledge, domain knowledge and the amount of time they invest. This inevitably hinders these methods’ ability to scale up to a large and diverse question bank.

Data-driven methods, deep neural network based methods, in particular, have recently emerged as a promising approach for various natural language processing tasks, such as machine translation, named entity recognition and sentiment classification. Inspired by the success of these works, Du et al. [51] treated question generation processes as a sequence-to-sequence learning problem, which directly maps a piece of text (usually a sentence) to a question. In contrast to rule-based methods, these methods can capture complex question generation patterns from data without handcrafted rules, thus being much more scalable than rule-based methods. As with most data-driven ap-
proaches, the success of neural network based methods is largely dependent on the size of the dataset as well as its quality [131].

<table>
<thead>
<tr>
<th>Source</th>
<th>Document-Question pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>Doc: ... after Heine’s German birthplace of Düsseldorf had rejected, allegedly for anti-Semitic motives ... Q: Where was Heine born?</td>
</tr>
<tr>
<td>RACE</td>
<td>Doc: ... There is a big supermarket near Mrs. Green’s home. She usually ... Q: Where is the supermarket?</td>
</tr>
<tr>
<td>LearningQ</td>
<td>Doc: ... gases have energy that is proportional to the temperature. The higher the temperature, the higher the energy the gases have. The crazy thing is that at the same temperature, all gases have the same energy ... Q: If you were given oxygen (molecular mass = 18 AMU) and hydrogen (molecular mass = 1 AMU) at the same temperature and pressure, which has more energy?</td>
</tr>
</tbody>
</table>

Table 6.1: Examples of document-question pairs.

Exiting datasets, such as SQuAD [131] and RACE [96], though containing a large number of questions (e.g., 97K questions in SQuAD), are not suitable for question generation in the learning context. Instead of being aimed at educational question generation, these datasets were originally collected for reading comprehension tasks. They are often limited in their coverage of topics—the questions in SQuAD for example, were generated by crowdworkers based on a limited number (536) of Wikipedia articles. More importantly, these questions seek factual details, and the answer to each question can be found as a piece of text in the source passages; they do not require higher-level cognitive skills to answer them, as exemplified by the SQuAD and RACE example questions in Table 6.1. We speculate, as a consequence, question generators built on these datasets cannot generate questions of varying cognitive skill levels and knowledge dimensions that require a substantial amount of cognitive efforts to answer, which unavoidably limits the applicability of the trained question generators for educational purpose.

To address these problems, we investigate the research question: RQ 5.1 can a large-scale and high-quality educational question dataset
be collected from the Social Web? Specifically, we present LearningQ, which consists of more than 230K document-question pairs collected from mainstream online learning platforms. LearningQ does not only contain questions designed by instructors (7K) but also questions generated by students (223K) during their learning processes, e.g., watching videos and reading recommended materials. It covers a diverse set of educational topics ranging from computing, science, business, humanities, to math. Through both quantitative and qualitative analyses, we show that, compared to existing datasets, LearningQ contains more diverse and complex source documents; moreover, solving the questions requires higher-order cognitive skills (e.g., applying, analyzing). Specifically, we show that most questions in LearningQ are relevant to multiple source sentences in the corresponding document, suggesting that effective question generation requires reasoning over the relationships between document sentences, as shown by the LearningQ question example in Table 6.1. Besides, we evaluate both rule-based and state-of-the-art deep neural network based question generation methods on LearningQ. Our results show that methods which perform well on existing datasets cannot generate high-quality educational questions, suggesting that LearningQ is a challenging dataset worth of significant further study.

Have you ever dropped your swimming goggles in the deepest part of the pool and tried to swim down to get them? It can be frustrating because the water tries to push you back up to the surface as you’re swimming downward. The name of this upward force exerted on objects submerged in fluids is called the buoyant force.

Table 6.2: Question-worthy sentence in a paragraph.

After presenting LearningQ, we further investigate the problem of identifying question-worthy content (i.e., sentences used as input for the question generator) from an article, which is largely ignored by existing studies \cite{51, 69}. Formally, we direct our efforts in answering the research question: RQ 5.2 what are effective strategies in identifying question-worthy sentences from an article? To be specific, given a paragraph or an article, often there are only a limited number of sentences that are worth asking questions about, i.e., those carrying important concepts. An example is shown in Table 6.2, where the last sentence defines the most important concept “buoyant force”. We, therefore, argue that selecting question-worthy sentences is of critical importance to the generation of high-quality educational questions.
To this end, we aim at achieving a better understanding of the effectiveness of different textual features in identifying question-worthy sentences from an article and proposing a total of nine strategies for question-worthy sentence selection, which cover a wide range of possible question-asking patterns inspired by both low-level and high-level textural features. For instance, we represent our assumption that informative sentences are more likely to be asked about by leveraging low-level features such as sentence length and the number of concepts as informativeness metrics; our assumption that important sentences are more worth asking about is represented by leveraging semantic relevance between sentences, which can be measured by using summative sentence identification techniques [53, 23]. To evaluate the effectiveness of the proposed strategies, we apply them to identify question-worthy sentences on five question generation datasets, i.e., SQuAD [13], TriviaQA [83], MCTest [133], RACE [96] and LearningQ presented in this chapter. We use the sentences identified by the proposed strategies as input for a well-trained question generator and evaluate the effectiveness of sentence selection strategies by comparing the quality of the generated questions.

To the best of our knowledge, LearningQ is the first large-scale dataset for educational question generation. It provides a valuable data source for studying cross-domain question generation patterns. The distinct features of LearningQ make it a challenging dataset for driving the advances of automatic question generation methods. Also, our work is the first to systematically study question-worthy sentence selection strategies across multiple datasets. Through extensive experiments, we find that the most question-worthy sentences in Wikipedia articles are often the beginning ones. In contrast, questions collected for learning purposes usually feature a more diverse set of sentences, including those that are most informative, important, or contain the largest amount of novel information. We further demonstrate that LexRank, which identifies important sentences by calculating their eigenvector centrality in the graph representation of sentence similarities, gives the most robust performance across different datasets among the nine selection strategies.

6.2 Related Work

6.2.1 Question Generation

Automatic question generation has been envisioned since the late 1960s [135]. It is generally believed by learning scientists that the generation of high-
Chapter 6. LearningQ for Educational Question Generation

Quality learning questions should be based on the foundations of linguistic knowledge and domain knowledge, and thus they typically approach the task in a rule-based manner \cite{161, 3, 69, 109}. Such rules mainly employ syntactic transformations to turn declarative sentences into interrogative questions \cite{35}. For instance, Mitkov and Ha \cite{109} generated multiple-choice questions from documents by employing rules of term extraction. Based on a set of manually-defined rules, Heilman and Smith \cite{69} produced questions in an overgenerate-and-rank manner where questions are ranked based on their linguistic features. These methods, however, are intrinsically limited in scalability: rules developed in certain subjects (e.g., introductory linguistics, English learning) cannot be easily adapted to other domains; the process of defining rules requires considerable efforts from experienced teachers or domain experts. More importantly, manually designed rules are often incomplete and do not cover all possible document-question transformation patterns, thus limiting rule-based generators to produce high-quality questions.

Entering the era of large-scale online learning, e.g., Massive Open Online Courses \cite{119}, the demand for automatic question generation has been increasing rapidly along with the largely increased number of learners and online courses accessible to them. To meet the need, more advanced computational techniques, e.g., deep neural network based methods, have been proposed by computer scientists \cite{51, 50}. In the pioneering work by Du et al. \cite{51}, an encoder-decoder sequence learning framework \cite{144} incorporated with the global attention mechanism \cite{103} was used for question generation. The proposed model can automatically capture question-asking patterns from the data, without relying on any hand-crafted rules, thus has achieved superior performance to rule-based methods regarding both scalability and the quality of the generated questions.

These methods, however, have only been tested on datasets that were originally collected for machine reading comprehension tasks. Noticeably, these datasets contain a very limited number of useful questions for learning, as we will show in the following sections. Therefore, it remains an open question how deep neural network methods perform in processing complex learning documents and generating desirable educational questions.

6.2.2 Datasets for Question Generation

Several large-scale datasets have been collected to fuel the development of machine reading comprehension models, including SQuAD \cite{131}, RACE \cite{96}, NewsQA \cite{150}, TriviaQA \cite{83}, NarrativeQA \cite{93}, etc. Though containing
questions, all of these datasets are not suitable for educational question generation due to either the limited number of topics \cite{31, 96} or the loose dependency between documents and questions, i.e., a document might not contain the content necessary to generate a question and further answer the question. More importantly, most questions in these datasets are not specifically designed for learning activities. For example, SQuAD questions were generated by online crowdworkers and are used to seek factual details in source documents; TriviaQA questions were retrieved from online trivia websites. An exception is RACE, which was collected from English examinations designed for middle school and high school students in China. Though collected in a learning context, RACE questions are mainly used to assess students’ knowledge level of English, instead of other skills or knowledge of diverse learning subjects.

Depending on different teaching activities and learning goals, educational questions are expected to vary in cognitive complexity, i.e., requiring different levels of cognitive efforts from learners to solve. Ideally, an educational question generator should be able to generate questions of all cognitive complexity levels, e.g., from low-order recalling factual details to high-order judging the value of a new idea. This requires the dataset for training educational question generators to contain questions of different cognitive levels. As will be presented in our analysis, LearningQ covers a wide spectrum of learning subjects as well as cognitive complexity levels and is therefore expected to drive forward the research on automatic question generation.

6.2.3 Question-worthy Sentence Selection

Existing studies, however, pay little attention to the selection of question-worthy sentences: they either assume that the question-worthy sentences have been identified already \cite{51} or simply take every sentence in an article as input for the question generator. For instance, \cite{69} assume that all sentences in an article are question-worthy and thus generate one question for each sentence and select high-quality ones based on their linguistic features. To our knowledge, \cite{50} is the only study that explicitly tackles the question-worthy sentence selection problem. It uses a bidirectional LSTM network \cite{75} to simultaneously encode a paragraph and calculate the question-worthiness of a sentence in the paragraph. However, training such a network relies on a large amount of ground-truth labels of question-worthy sentences (e.g., tens of thousands). Obtaining these labels is a long, laborious, and usually costly process. Furthermore, the proposed deep neural network was only validated in short paragraphs instead of the whole article. Considering that reading
materials can be much longer and deep neural networks can fail at processing long sequence data due to the vanishing gradient problem \([10]\), it remains an open question whether the proposed method can handle long articles.

Instead of developing a novel neural network architecture that simultaneously does sentence selection and question generation (like \([51]\) does), we take one step back and focus extensively on question-worthy sentence selection. We propose heuristic strategies which exploit different textual features in selecting question-worthy sentences from an article, so as to clarify the main criteria in the selection process, and to adequately inform educational question generator design.

### 6.3 Data Collection

#### 6.3.1 Data Sources

To gather large amounts of useful learning questions, we initially explored several mainstream online learning platforms and finally settled on two after having considered the data accessibility and the quantity of the available questions as well as the corresponding source documents. Concretely, we gathered LearningQ data from the following two platforms:

1. **TED-Ed**\(^1\) is an education initiative supported by TED which aims to spread the ideas and knowledge of teachers and students around the world. In TED-Ed, teachers can create their own interactive lessons, which usually involve lecture videos along with a set of carefully crafted questions to assess students’ knowledge. Lesson topics range from humanity subjects like arts, language, and philosophy to science subjects like business, economics and computer technology. Typically, a lesson, covering a single topic, includes one lecture video, and lasts from 5 to 15 minutes. Due to the subscription-free availability, TED-Ed has grown into one of the most popular educational communities and served millions of teachers and students every week. As questions in TED-Ed are created by instructors, we consider them to be high-quality representations of testing cognitive skills at various levels (e.g., the LearningQ question in Table \(6.1\) is from TED-Ed). We use TED-Ed as the major data source to collect instructor-crafted learning questions.

2. **Khan Academy**\(^2\) is another popular online learning platform. Similar to TED-Ed, Khan Academy also offers lessons to students around the world. Compared to

---

1. https://ed.ted.com/
2. https://www.khanacademy.org/
6.3. Data Collection

TED-Ed, the lessons are targeted at a wider audience. For example, the math subjects in Khan Academy cover topics from kindergarten to high school. In addition, the lessons are organized in alignment with typical school curriculum (from the easier to the more advanced) instead of being an independent collection of videos as is the case in TED-Ed. Another distinction between the two platforms is that Khan Academy allows learners to leave posts and ask questions about the learning materials (i.e., lecture videos and reading materials) during their learning. For instance, the chemistry course Quantum numbers and orbitals\(^3\) includes one article (titled The quantum mechanical model of the atom) and three lecture videos (titled Heisenberg uncertainty principle, Quantum numbers and Quantum numbers for the first four shells) and learners can ask questions about any of them. More often than not, learners’ questions express their confusion about the learning material—e.g., “How do you convert Celsius to Calvin?”—and thus are an expression of learners’ knowledge gaps that need to be overcome to master the learning material. We argue that these questions can promote in-depth thinking and discussion among learners, thus complementing instructor-designed questions. We use those learner-generated questions as part of LearningQ.

We implemented site-specific crawlers for both Khan Academy and TED-Ed and collected all available questions and posts in English as well as their source documents at both platforms that were posted on or before December 31, 2017, resulting in a total of 1,146,299 questions and posts.

6.3.2 Question Classification for Khan Academy

Compared to instructor-designed questions collected from TED-Ed, learner-generated posts in Khan Academy can be of relatively low quality for our purposes since they are not guaranteed to contain a question (a learner may for example simply express her appreciation for the video), or the contained question can be off-topic, lack the proper context, or be too generic. Examples of high- and low-quality questions are shown in Table 6.3.

Originally, we gathered a total of 953,998 posts related to lecture videos and 192,301 posts related to articles from Khan Academy. To distill useful learning questions from the collected posts, we first extracted sentences ending with a question mark from all of the posts, which resulted in 407,723 such sentences from posts on lecture videos and 66,100 on reading material. To further discriminate useful questions for learning from non-useful ones, we ran-

domly sampled 5,600 of these questions and recruited two education experts to annotate the questions: each expert labeled 3,100 questions (600 questions were labeled by both experts to determine the inter-annotator agreement) in a binary fashion: useful for learning or not. Based on the labeled data, we trained a convolutional neural network on top of pre-trained word embeddings to classify the remaining Khan Academy questions. In the following, we describe the labeling process in more details.

<table>
<thead>
<tr>
<th>ID</th>
<th>Questions</th>
<th>Topic</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>a)</td>
<td>What is the direction of current in a circuit?</td>
<td>S</td>
<td>√</td>
</tr>
<tr>
<td>b)</td>
<td>Why can’t voltage-gated channels be placed on the surface of Myelin?</td>
<td>S</td>
<td>√</td>
</tr>
<tr>
<td>c)</td>
<td>Is there a badge for finishing this course?</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>d)</td>
<td>Have you looked on your badges page to see if it is one of the available badges?</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>e)</td>
<td>Why do each of them have navels?</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>f)</td>
<td>Does it represent phase difference between resistance and reactance?</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>g)</td>
<td>What should the graph look like for higher voltages?</td>
<td>S</td>
<td>√</td>
</tr>
<tr>
<td>h)</td>
<td>What if some of the ideas come from different historical perspectives, giving inaccurate information?</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>i)</td>
<td>What if the information is wrong?</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>j)</td>
<td>Can someone please help me?</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>k)</td>
<td>Could you be more specific?</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>l)</td>
<td>Are you asking what geometric means?</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>m)</td>
<td>Are you talking about the frequency?</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>n)</td>
<td>What programming language or how much of coding I need to know to start learning algorithms here?</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>o)</td>
<td>Can I do algorithms or should I do programming first?</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Examples of useful (marked with √) and non-useful questions from Khan Academy. S/H/M/C/E/T denote Science, Humanities, Math, Computing, Economics and Test Preparation, respectively.
6.3. Data Collection

**Question Annotation.** We consider a user-generated question to be as useful for learning when all of the following conditions hold: (i) the question is *concept-relevant*, i.e., it seeks for information on knowledge concepts taught in lecture videos or articles; (ii) the question is *context-complete*, which means sufficient context information is provided to enable other learners to answer the question; and (iii) the question is not generic (e.g., a question asks for learning advice). To exemplify this, two concept-relevant learning questions are shown in Table 6.3 (a and b), accompanied by two concept-irrelevant ones (c and d). Question e and f in the same table are also concept-relevant. However, as they don’t provide enough context information, e.g., lack of references for “they” and “it”, we consider them as non-useful. As a counter-example, we consider question g in the table as useful since the reference for “the graph” can be easily inferred. This comes in contrast to question h and i, where the references for “the idea” and “the information” are too vague thus failed to provide sufficient context information. Finally, generic questions expressing the need for help (j and k), asking for clarification (l and m) or general learning advice (as exemplified by n and o), are not useful for learning the specific knowledge concepts.

**Annotation & Classification Results.** Of the 5,600 annotated questions, we found 3,465 (61.9%) to be useful questions for learning. The inter-annotator agreement reached a Cohen’s Kappa of 0.82, which suggests a substantially coherent perception of question usefulness by the two annotators. To understand the performance of the classifier trained on this labeled dataset, we randomly split the dataset into a training set of size 5,000, a validation set of size 300, and a test set of size 300. We iterated the training and evaluation process 20 times to obtain a reliable estimation of classification performance. Results show that the model reaches an accuracy of 80.5% on average (SD=1.8%), suggesting that the classifier can be confidently applied for useful/non-useful question classification. With this classifier, we retain about 223K unique useful questions in Khan Academy, which will be used for our following analysis.

### 6.3.3 Final Statistics of LearningQ

An overall description of LearningQ is shown in Table 6.4 (rows 1—4). As a means of comparison, we also provide the same statistics for the popular question generation datasets (though not necessarily useful for education and learning) SQuAD and RACE. Compared to these two datasets, LearningQ (i) consists of about 230K questions (versus 97K in SQuAD and 72K in RACE) on nearly 11K source documents; (ii) contains not only useful educational
<table>
<thead>
<tr>
<th>Row</th>
<th>Feature Type</th>
<th>Features</th>
<th>SQuAD</th>
<th>RACE</th>
<th>TED-Ed</th>
<th>Khan Academy Video</th>
<th>Khan Academy Learner</th>
<th>Khan Academy Learner</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Creator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Basic statistics</td>
<td># Unique documents</td>
<td>20,958</td>
<td>27,933</td>
<td>1,102</td>
<td>7,924</td>
<td>1,815</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Basic statistics</td>
<td># Unique questions</td>
<td>97,888</td>
<td>72,547</td>
<td>7,612</td>
<td>201,273</td>
<td>22,585</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Basic statistics</td>
<td># Avg. questions / document</td>
<td>4.67</td>
<td>2.60</td>
<td>6.91</td>
<td>25.40</td>
<td>12.44</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Document &amp; question length</td>
<td># Avg. words / document</td>
<td>134.84</td>
<td>322.88</td>
<td>847.64</td>
<td>1370.83</td>
<td>1306.55</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Document &amp; question length</td>
<td># Avg. sentence / document</td>
<td>4.96</td>
<td>17.63</td>
<td>42.89</td>
<td>73.51</td>
<td>49.85</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Document &amp; question length</td>
<td># Avg. words / sentence of documents</td>
<td>27.17</td>
<td>18.31</td>
<td>19.76</td>
<td>18.65</td>
<td>26.21</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Document &amp; question length</td>
<td># Avg. words / question</td>
<td>11.31</td>
<td>11.51</td>
<td>20.07</td>
<td>16.72</td>
<td>17.11</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Document &amp; question length</td>
<td># Avg. sentence / question</td>
<td>1.00</td>
<td>1.03</td>
<td>1.41</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Entities</td>
<td># Avg. entities /document</td>
<td>10.24</td>
<td>9.75</td>
<td>17.66</td>
<td>14.55</td>
<td>47.38</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Entities</td>
<td># Avg. entities /question</td>
<td>0.92</td>
<td>0.53</td>
<td>0.66</td>
<td>0.29</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Entities</td>
<td>% Entity words in question</td>
<td>8.10</td>
<td>4.58</td>
<td>3.29</td>
<td>1.72</td>
<td>2.54</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Readability</td>
<td>Document readability</td>
<td>45.82</td>
<td>73.49</td>
<td>64.08</td>
<td>76.54</td>
<td>55.15</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Readability</td>
<td>Question readability</td>
<td>67.23</td>
<td>51.00</td>
<td>66.32</td>
<td>72.15</td>
<td>69.04</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: Descriptive features and statistics of LearningQ and the datasets in comparison.
questions carefully designed by instructors but also those generated by learners for in-depth understanding and further discussion of the learning subject; (iii) covers a wide range of educational subjects from two mainstream online learning platforms. To highlight the characteristics of LearningQ, we also include SQuAD and RACE in the data analysis presented next.

6.4 Sentence Selection Strategies

In the following, we describe in detail our proposed sentence selection strategies based on different question-asking assumptions and sentence properties measured by different textual features.

• **Random sentence (Random).** As the baseline, we randomly select a sentence and use it as input for the question generator.

• **Longest sentence (Longest).** This strategy selects the longest sentence in an article. The assumption is that people tend to ask questions about sentences containing a large amount of information, which, intuitively, can be measured by their lengths.

• **Concept-rich sentence (Concept).** Different from Longest, this strategy assumes that the amount of information can be better measured by the total number of entities in a sentence. The more entities a sentence contains, the richer information it has.

• **Concept-type-rich sentence (ConceptType).** This strategy is a variant of Concept. It calculates the total number of entity types in a sentence to measure the informativeness of a sentence.

The above three strategies approximate question-worthiness of a sentence by *informativeness*, which is further measured by different textual features. In contrast, the following two strategies approximate question-worthiness of a sentence by *difficulty* and *novelty*, respectively.

• **Most difficult sentence (Hardest).** This strategy is built on the assumption that difficult sentences can sometimes bring the most important messages that should be questioned and assessed. Therefore, it chooses the most difficult sentence in an article as the question-worthy sentence. We calculate the Flesch Reading Ease Score \[32\] of sentences as their difficulty indicators.
Chapter 6. LearningQ for Educational Question Generation

- **Novel sentence (Novel).** Unlike Hardest, this strategy believes that sentences with novel information that people do not know before are more question-worthy. We calculate the number of words that never appear in previous sentences as a sentence’s novelty score and select the most novel one.

Finally, we introduce three strategies that approximate question-worthiness of a sentence by the relative importance of the sentence concerning the remaining ones in an article. The importance is either measured by the relative position of a sentence or its centrality represented by semantic relevance with other sentences.

- **Beginning sentence (Beginning).** In the research of text summarization, one common hypothesis about sentence positions is the importance of a sentence decreases with its distance from the beginning of the document, and therefore less question-worthy. This strategy selects the first sentence in an article as the most question-worthy sentence.

- **Centroid based important sentence (LexRank).** In line with Beginning, this strategy also believes that question-worthy sentences should be selected from those of greater importance. The difference here is that the sentence importance is measured by the centroid-based method, LexRank, which calculates sentence importance based on eigenvector centrality in a graph of sentence similarities.

- **Maximum marginal relevance based important sentence (MMR).** Different from LexRank, this strategy computes sentence importance by considering a linear trade-off between relevance and redundancy, i.e., selecting the sentence that is most relevant but shares least similarity with the other sentences as the most important sentence.

6.5 Data Analysis on LearningQ

The complexity of questions concerning the required cognitive skill levels and knowledge dimensions is a crucial property that can significantly influence the quality of questions for learning. We thus believe that this factor should be studied when building efficient question generators. However, to
our knowledge, there is no work attempting to characterize this property of questions in datasets for question generation.

In this section, we characterize the cognitive complexity of questions in LearningQ and other existing question generation datasets (i.e., SQuAD and RACE as two representatives) along several dimensions: (i) low-level document and question attributes related to cognitive complexity [167, 170], e.g., the number of sentences or words per document or per question; (ii) document and question properties that can affect human perception of cognitive complexity, which include topical diversity, document and question readability [68, 147], etc.; and (iii) cognitive skill levels in accordance with Bloom’s Revised Taxonomy [104].

6.5.1 Document & Question Lengths

Table 6.4 (rows 5—9) presents statistics on document and question lengths. It can be observed that while, on average, the number of words per sentence in the documents of LearningQ are not larger than in SQuAD/RACE, documents from both TED-Ed and Khan Academy are more than twice as longer than those from SQuAD and RACE. In particular, SQuAD documents are on average nearly ten times shorter than Khan Academy documents. The same observation holds for the questions in LearningQ, where question length is twice as long as that of SQuAD and RACE. Compared with those in Khan Academy, documents in TED-Ed are shorter. This is mainly due to the fact that TED-Ed encourages shorter videos on a single topic.

6.5.2 Topics, Interrogative Words, and Readability

To gain an overview of the topics, we applied Named Entity Recognition to obtain statistics on the entities. The results are shown in rows 10 and 11 of Table 6.4. To gain more insights into the semantics of the documents and questions, we report the most frequent terms (after stopword removal) in Table 6.5 across both documents and questions. To gain insights into the type of questions, we separately consider interrogative terms (such as who or why) in the rightmost part of Table 6.5 by keeping most stopwords but filtering out prepositions and definite articles.

We observe in Table 6.4 that documents in LearningQ on average contains 160% more entities than SQuAD and RACE, which is expected because LearningQ documents are longer. Yet, the number of entities in LearningQ questions are not significantly larger than SQuAD and RACE. In particular,
Table 6.5: Top words in documents and questions and top interrogative words of questions in LearningQ and the datasets in comparison. Words pertinent to a specific data source platforms are in bold. KA represents Khan Academy.
questions in SQuAD contain 40% more entities than those in LearningQ. This is despite the fact that SQuAD documents are shortest overall, as we showed earlier. To eliminate the influence of question lengths and refine the analysis, we further observe that the percentage of entities among all the words in SQuAD questions is higher than that in LearningQ questions. The same observation holds when comparing RACE with LearningQ. These observations imply that, on the one hand, documents in LearningQ are more complex concerning the number of involved entities; on the other hand, fewer questions related to entities, i.e., fewer factual questions, exist in LearningQ than the other datasets.

This interpretation is also supported by the top-k words shown in Table 6.5. We observe that while both documents and questions in SQuAD favor topics related to time and location (e.g., time, year, century, city, state), all data sources in LearningQ have fewer questions on these topics; more often in LearningQ questions, we find abstract words such as mean, difference, function, which are indicative of higher level cognitive skills being required. In line with this observation, we note that more interrogative words seeking factual details such as who and when rank high in the list of starting words of questions in SQuAD, while questions in LearningQ sources start much more frequently with why. This suggests that answering LearningQ questions often requires a deeper understanding of learning materials. Interestingly, one can observe in TED-Ed questions (in the middle part of the table) frequent words such as think and explain, which explicitly ask learners to process learning materials in a specific way. These required actions are directly related to learning objectives defined by Bloom’s Revised Taxonomy, as we will analyze later. In addition to the above, another interesting observation from Table 6.5 is that learners frequently ask questions for the clarification of videos using words such as know, Sal (the name of the instructor who initially created most videos at Khan Academy in the early stage of the platform), and mean.

Readability. Readability is an important document and question property related to learning performance. Table 6.4 (rows 13—14) reports the Flesch readability scores of documents and questions in the compared datasets [39]. A piece of text with larger a Flesch readability score indicates it is easier to understand. Questions found alongside both Khan Academy videos and article possess similar readability scores, despite the different sources. This confirms our previous finding on the similarity between the two subsets of Khan Academy data. We, therefore, do not distinguish these two subsets in the following analyses focused on questions.
6.5.3 Cognitive Skill Levels

It is generally accepted in educational research that a good performance on assessment questions usually translates into “good learning” [72]. We first use Bloom’s Revised Taxonomy to categorize the questions according to the required cognitive efforts behind them [104]. The taxonomy provides guidelines for educators to write learning objectives, design the curriculum and create assessment items aligned with the learning objectives. It consists of six learning objectives (requiring different cognitive skill levels from lower order to higher order):

- **Remembering**: questions that are designed for retrieving relevant knowledge from long-term memory.
- **Understanding**: questions that require constructing meaning from instructional messages, including oral, written and graphic communication.
- **Applying**: questions that ask for carrying out or using a procedure in a given situation.
- **Analyzing**: questions that require learners to break material into constituent parts and determine how parts relate to one another and to an overall structure or purpose.
- **Evaluating**: questions that ask for make judgments based on criteria and standards.
- **Creating**: questions that require learners to put elements together to form a coherent whole or to re-organize into a new pattern or structure.

To exemplify, we select one question example for each category that we collected from TED-Ed and Khan Academy, as shown in Table 6.6. Among the different learning objectives defined by Bloom’s Revised Taxonomy, analyzing is an objective closely related to the task of automatic question generation. *Analyzing* questions require the learner to understand the relationships between different parts of the learning material. Existing question generation methods [51], however, can usually only take one sentence as input. To cope with *analyzing* questions, state-of-the-art methods first need to determine the most relevant sentence in the learning material, which is then used as input to the question generator. This inevitably limits the ability of trained question generators to deliver meaningful *analyzing* questions covering multiple
<table>
<thead>
<tr>
<th>Taxonomy</th>
<th>TED-Ed Examples</th>
<th>Khan Academy Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remembering</td>
<td>How big is an atom?</td>
<td>What is a negative and a positive feedback in homeostasis?</td>
</tr>
<tr>
<td>Understanding</td>
<td>Why do some plankton migrate vertically?</td>
<td>Why can’t voltage-gated channels be placed on the surface of myelin?</td>
</tr>
<tr>
<td>Applying</td>
<td>What kind of invention would you make with shape memory materials if you could get it in any form you wanted?</td>
<td>If i double the area and take the half of the fraction, do I get the same result?</td>
</tr>
<tr>
<td>Analyzing</td>
<td>Why are cities like London, Tokyo, and New York facing shortages in burial ground space?</td>
<td>Why did sea levels drop during the ice age?</td>
</tr>
<tr>
<td>Evaluating</td>
<td>Mansa Musa is one of many African monarchs throughout the continent’s rich history. Yet, the narratives of only a few kings and queens are featured in television and movies. Analyze and evaluate why you think that this is the case, then create two ideas for how we can work to bring more positive awareness of the history of Africa’s ancient and contemporary kings and queens to students today.</td>
<td>Will all the cultures merge into one big culture, due to the fading genetic distinctions?</td>
</tr>
<tr>
<td>Creating</td>
<td>Can somebody please explain to me what marginal benefits is and give me some examples?</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.6: Question Examples of Different Bloom’ Revised Taxonomy Level in TED-Ed and Khan Academy.
knowledge concepts scattered in the source documents. To understand the complexity of the LearningQ questions specifically from the point of view of training question generators, we also include in our analyses an exploration of the proportion of questions at various Bloom levels that require knowledge from multiple source document sentences.

**Data Annotation.** To facilitate our analysis, we recruited two experienced instructors to label 200 randomly sampled questions from each of the compared datasets according to Bloom’s Revised Taxonomy. The Cohen’s Kappa agreement score between the two annotators reached 0.73, which is a substantial agreement. In a second labeling step, we labeled the selected questions with their sentence(s) based on which they are generated.

**Comparative Results.** Table 6.7 shows the results of question classification according to Bloom’s Revised Taxonomy. SQuAD only contains remembering questions, suggesting that it is the least complex dataset among all compared datasets regarding required cognitive skill levels. In general, we note a trend of decreasing percent of remembering questions (and increasing percentage of understanding questions) from SQuAD, RACE, to TED-Ed and Khan Academy. We can conclude that questions in LearningQ demand higher cognitive skills than those in SQuAD and RACE. Interestingly, among the two different LearningQ sources, we can observe that there are more understanding and applying questions in Khan Academy than in TED-Ed, while there are more evaluating and creating questions in TED-Ed than in Khan Academy. This shows the inherent differences related to the corresponding learning activities between instructor-designed questions and learner-generated questions. The former is mainly used for assessment purpose and thus contains more questions of higher-order cognitive skill levels; the latter is generated during students’ learning process (e.g., watching lecture videos and reading recommended materials) and is usually used to seek for a better understanding of the learning material. Note that 26.42% Khan Academy questions were labeled as either irrelevant or unknown due to being not useful for learning or missing enough context information for the labeler to assign a Bloom category. This aligns with the accuracy of the useful question classifier we reported in the data collection section.

In Table 6.8 we report the results of our source sentence(s) labeling efforts. From the statistics of # words in source sentences, we can observe an increasing requirement for reasoning over multiple sentences from SQuAD and RACE to TED-Ed and Khan Academy. Compared to the 98.5% of single sentence related questions in SQuAD, questions in TED-Ed (Khan Academy) are related to 3.53 (6.65) sentences on average in source documents. In par-
6.6. Experiments and Results

In this section, we first conduct experiments to evaluate the performance of rule-based and deep neural network based methods in question generation using LearningQ. We aim to answer the following questions: 1) how effective are these methods at generating high-quality educational questions; 2) to what extent is their performance influenced by the learning topics; and 3) to

Table 6.7: Distribution of Bloom’s Revised Taxonomy Labels.

<table>
<thead>
<tr>
<th>taxonomy</th>
<th>SQuAD</th>
<th>RACE</th>
<th>TED-Ed</th>
<th>Khan Academy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remembering</td>
<td>100</td>
<td>82.19</td>
<td>61.86</td>
<td>18.24</td>
</tr>
<tr>
<td>Understanding</td>
<td>0</td>
<td>18.26</td>
<td>38.66</td>
<td>55.97</td>
</tr>
<tr>
<td>Applying</td>
<td>0</td>
<td>0.46</td>
<td>9.79</td>
<td>12.58</td>
</tr>
<tr>
<td>Analyzing</td>
<td>0</td>
<td>8.22</td>
<td>14.95</td>
<td>15.09</td>
</tr>
<tr>
<td>Evaluating</td>
<td>0</td>
<td>1.37</td>
<td>4.12</td>
<td>1.89</td>
</tr>
<tr>
<td>Creating</td>
<td>0</td>
<td>0</td>
<td>1.55</td>
<td>0.63</td>
</tr>
<tr>
<td>Unknown/Irrelev</td>
<td>0</td>
<td>3.20</td>
<td>0</td>
<td>26.42</td>
</tr>
</tbody>
</table>

Table 6.8: Results of Source Sentence Labelling. # Words/Sent. denote the average words/sentences in the labelled source sentences. % ONE/MULTIPLE/EXTERNAL refer to the percentage of questions related to ONE single sentence, MULTIPLE sentences or require EXTERNAL knowledge to generate, respectively. KA denotes Khan Academy.

<table>
<thead>
<tr>
<th></th>
<th>SQuAD</th>
<th>RACE</th>
<th>TED-Ed</th>
<th>KA</th>
</tr>
</thead>
<tbody>
<tr>
<td># Words</td>
<td>32.39</td>
<td>46.02</td>
<td>76.57</td>
<td>128.23</td>
</tr>
<tr>
<td># Sent.</td>
<td>1.01</td>
<td>2.87</td>
<td>3.53</td>
<td>6.65</td>
</tr>
<tr>
<td>% ONE</td>
<td>98.53</td>
<td>37.10</td>
<td>28.63</td>
<td>9.43</td>
</tr>
<tr>
<td>% MULTIPLE</td>
<td>1.47</td>
<td>62.90</td>
<td>52.42</td>
<td>23.27</td>
</tr>
<tr>
<td>% EXTERNAL</td>
<td>0</td>
<td>0</td>
<td>18.95</td>
<td>38.99</td>
</tr>
</tbody>
</table>

In particular, Table 6.8 (the last row) further shows that a large portion of the questions in LearningQ, especially in Khan Academy, cannot be answered by simply relying on the source document, as exemplified by the evaluating/creating question from TED-Ed in Table 6.6 and thus require external knowledge to generate.

6.6 Experiments and Results

In this section, we first conduct experiments to evaluate the performance of rule-based and deep neural network based methods in question generation using LearningQ. We aim to answer the following questions: 1) how effective are these methods at generating high-quality educational questions; 2) to what extent is their performance influenced by the learning topics; and 3) to
Chapter 6. LearningQ for Educational Question Generation

what extent does the source sentence(s) length affect the question generation performance.

Then, we evaluate the performance of the strategies for question-worthy sentence selection as proposed in Section 6.4 across the five datasets.

6.6.1 Experimental Setup

Comparison Methods for LearningQ Evaluation. We investigate a representative rule-based baseline and two state-of-the-art deep neural networks in question generation:

- **H&S** is a rule-based system which can be used to generate questions from source text for educational assessment and practice [69]. The system produces questions in an overgenerate-and-rank manner. We only evaluate the top-ranked question.

- **Seq2Seq** is a representative encoder-decoder sequence learning framework proposed for machine translation [144]. It automatically learns the patterns of transforming an input sentence into an output sentence based on training data.

- **Attention Seq2Seq** is the state-of-the-art method proposed in [51], which incorporates the global attention mechanism [103] into the encoder-decoder sequence learning framework during the decoding process. The attention mechanism allows the model to mimic humans problem-solving process by focusing on relevant information in the source text and using this information to generate a question.

We implemented the two neural network based methods on top of the OpenNMT system [92]. In accordance with the original work [144, 51], Bi-LSTM is used for the encoder and LSTM for the decoder. We tune all hyperparameters using the held-out validation set and select the parameters that achieve the lowest perplexity on the validation set. The number of LSTM layers is set to 2, and its number of hidden units is set to 600. The dimension of input word embedding is set to 300 and we use the pre-trained embeddings glove.840B.300d for initialization [124]. Model optimization is performed by applying Adam [89]; we set the learning rate to 0.001 and the dropout rate to 0.3. The gradient is clipped if it exceeds 5. We train the models for 15 epochs in mini-batches of 64. When generating a question, beam search with a beam size of 3 is used, and the generation stops when every beam in the stack produces the <EOS> (end-of-sentence) token.
LearningQ Preparation. We use the NLTK tool \cite{16} to pre-process LearningQ: lower-casing, tokenization and sentence splitting. To account for the fact that existing methods can only process a small number of sentences as input, we need to decide the source sentences that each question corresponds to before the evaluation. Instead of applying the nine sentence selection strategies we propose in Section 6.4, which are evaluated and compared in Section 6.6.3, here for each question, we use the following strategy inspired by approaches for text similarity \cite{60} to locate the source sentences in the corresponding document most relevant to the question. If the target question contains a timestamp—e.g., “in 10:32, what does the Sal mean ...”—indicating the source sentence(s) location from which the target question is generated, we then choose that sentence as the starting sentence and compute the cosine similarity with the target question. We then go forwards and backwards in turns to determine whether including a nearby sentence would increase the cosine similarity between the target question and the source sentences. If yes, the nearby sentence is added. Otherwise, the search process stops. If a target question does not contain timestamp information, we select the sentence with the largest cosine similarity to the question to start our search the same way as described above to locate the source sentences. Due to the vanishing gradient problem in recurrent neural networks \cite{76}, we only keep data with source sentences containing no more than 100 words.

Notice that deep neural network based methods usually require a substantial amount of training data. The quantity of instructor-crafted questions in TED-Ed is not sufficient (7K). We, therefore, train the selected methods only on learner-generated questions. Concretely, we first merge all of the questions posted by Khan Academy learners on both lecture videos and reading materials, then randomly select 80% for training, 10% for validation and 10% for testing. At the same time, we also use all of the instructor-crafted questions as a second test set to investigate how effective the models built on learner-generated questions are in delivering instructor-crafted questions.

Unified Question Generator for Sentence Selection Evaluation. We also use Attention Seq2Seq as our testbed to evaluate the effectiveness of the proposed sentence selection strategies. To our knowledge, SQuAD is the only reading comprehension dataset with ground-truth labels of question-worthy sentence labels. Therefore, we use the labeled input sentences and the corresponding questions in SQuAD for training the question generator. We set the hyper-parameters as suggested in \cite{51}.

Notice that articles can be of different lengths and thus possibly contain different numbers of question-worthy sentences. During experiments, we be-
lieve the number of selected sentences should be dependent on the number of ground-truth questions gathered about an article: different ground-truth questions are seeking for different details about the article, i.e., based on different question-worthy sentences. We, therefore, evaluate the each of the questions generated by different selected sentences against all the ground-truth questions of the article and consider the result with the best performance as an indication of the selected sentence matched with the ground-truth question.

Datasets for Sentence Selection Evaluation. Generally, all reading comprehension datasets, i.e., those with questions and the corresponding documents which the questions are about, can be used to evaluate the selection strategies. We expect that the generated questions should be useful for learning purposes. Therefore, we select experimental datasets that contain natural questions designed by humans instead of search queries \cite{114} or cloze-style questions \cite{70}. With such consideration, we include five datasets for experiments: SQuAD \cite{131}, TriviaQA \cite{83}, MCTest \cite{133}, RACE \cite{96}, and LearningQ \cite{33}. TriviaQA contains questions from trivia and quiz-league websites and evidence documents gathered from web search and Wikipedia. Here we only consider questions with evidence documents collected from Wikipedia, which results in 138K question-document pairs. MCTest consists of 660 stories written by crowd-workers and 2K associated questions about the stories. Recall that LearningQ contains both instructor-designed questions gathered from TED-Ed and learner-generated questions gathered from Khan Academy. As the learner-generated questions can be redundant about the same knowledge concepts (i.e., same sentences), to avoid concept bias, we only include the 7K instructor-designed questions for experiments on sentence selection.

Questions in SQuAD and TriviaQA mainly seek for factual details and the answers can be found as a piece of text in the source paragraph/article from Wikipedia. Questions in MCTest are designed for young children. RACE and LearningQ are collected in learning contexts: RACE questions are mainly used to assess students’ knowledge level of English, whereas LearningQ covers a diverse set of educational topics, more complex articles, and the questions require higher-order cognitive skills to solve.

Evaluation Metrics. Similar to \cite{51}, we adopt Bleu 1, Bleu 2, Bleu 3, Bleu 4, Meteor and RougeL for evaluation. Bleu-n scores rely on the maximum n-grams for counting the co-occurrences between a generated question and a set of reference questions; the average of Bleu is employed as final score \cite{118}. Meteor computes the similarity between the generated question and the reference questions by taking synonyms, stemming and paraphrases into account.
6.6. Experiments and Results

\textit{Rouge}_L reports the recall rate of the generated question concerning the reference questions based on the longest common sub-sequence \cite{10}.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Bleu 4</th>
<th>Meteor</th>
<th>Rouge_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khan Academy</td>
<td>H&amp;S</td>
<td>0.28</td>
<td>0.17</td>
<td>0.13</td>
<td>0.10</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>Seq2Seq</td>
<td>19.84</td>
<td>7.68</td>
<td>4.02</td>
<td>2.29</td>
<td>6.44</td>
</tr>
<tr>
<td></td>
<td>Attn. Seq2Seq</td>
<td>24.70</td>
<td>11.68</td>
<td>6.36</td>
<td>3.63</td>
<td>8.73</td>
</tr>
<tr>
<td>TED-Ed</td>
<td>H&amp;S</td>
<td>0.38</td>
<td>0.22</td>
<td>0.17</td>
<td>0.15</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>Seq2Seq</td>
<td>12.96</td>
<td>3.95</td>
<td>1.82</td>
<td>0.73</td>
<td>4.34</td>
</tr>
<tr>
<td></td>
<td>Attn. Seq2Seq</td>
<td>15.83</td>
<td>5.63</td>
<td>2.63</td>
<td>1.15</td>
<td>5.32</td>
</tr>
</tbody>
</table>

Table 6.9: Performance of rule-based and deep neural network based methods on \textit{LearningQ}.

6.6.2 Evaluation on \textit{LearningQ}

\textbf{Results}. Table 6.9 reports the performance of the selected methods on learner-generated questions from \textit{Khan Academy} and instructor-designed questions from \textit{TED-Ed}. We can observe that across all different evaluation metrics, the rule-based method H&S is outperformed by both deep neural network based methods. This confirms previous findings in the new context of learning that data-driven methods are a better approach for question generation. Among the two deep neural network based methods, Attention Seq2Seq consistently outperform Seq2Seq ($p$-value < .001, Paired t-test). This verifies that the attention mechanism is an effective approach for boosting the performance of educational question generation.

By comparing the performance of the selected methods on \textit{Khan Academy} and on \textit{TED-Ed}, we find that the performance of rule-based method H&S varies across different evaluation metrics. The performance measured by Bleu scores are higher on learner-generated questions than on instructor-designed questions, while it is low as measured by Meteor and \textit{Rouge}_L. On the other hand, deep neural network based methods consistently reach a higher performance on learner-generated questions than on instructor-designed questions. Considering the fact that recurrent networks are less effective in handling long sentences, this could be due to two reasons: 1) the majority of questions in \textit{TED-Ed} are related to multiple sentences as we found (Table 6.8); and 2) the questions generated by learners are generally shorter than those designed by instructors (Table 6.4). In later analysis, we further describe how the length of source sentences would affect question generation performance.
Chapter 6. LearningQ for Educational Question Generation

The performance of the state-of-the-art methods is much lower on LearningQ than on existing datasets. Attention Seq2Seq achieves a Bleu 4 score $> 12$ and a Meteor score $> 16$ on SQuAD, while on LearningQ it only achieves Bleu 4 scores of $< 4/2$ and Meteor scores of $< 9/6$ on learner-generated questions/instructor-designed questions, respectively. Similar results also hold for the other metrics.

![Figure 6.1: Results of question generation on different learning subjects in Khan Academy.](image)

**Impacts of Subjects and Source Sentence Lengths.** We now investigate the performance of Attention Seq2Seq in generating educational questions as affected by different subjects and different lengths of input source sentences.

The impact of the source document topic on question generation performance is shown for Khan Academy in Figure 6.1. We observe that question generation performance varies across subjects. In particular, Bleu 4 varies from $< 2$ to $> 5$ for learner-generated questions and from 0.38 to 0.92 for instructor-designed questions. Compared to Economics and College Admissions, question generation for Math and Science can usually achieve higher performance. A similar variation is also observed on TED-Ed. These results indicate that topics can affect question generation performance. Fully understanding the co-influence of topics and other document properties (e.g., difficulty) however requires more studies, which we leave to future work.

As we showed before (Table 6.8), educational questions are related to multiple source sentences in the documents. However, existing neural network methods usually take only one or two source sentences as input to generate questions. To further investigate the effectiveness of existing methods when taking source sentences of different lengths as input, we divide the testing set according to the length of their source sentences. The results are shown in Figure 6.2. In general, question generation performance decreases when the length of source sentences increases across all metrics for both Khan Academy
6.6. Experiments and Results

and TED-Ed. This strongly suggests that the performance of the state-of-the-art method is significantly limited by long source sentences.

![Histograms showing performance metrics for different source sentence lengths](image)

(a) Khan Academy

![Histograms showing performance metrics for different source sentence lengths](image)

(b) TED-Ed

Figure 6.2: Results of question generation with different source sentence lengths.

6.6.3 Evaluation on Sentence Selection Strategies

Table 6.10 reports the results of the proposed sentence selection strategies on four datasets. We do not show results on the SQuAD dataset, as we found that the performance of different sentence selection strategies on SQuAD shows little variance, owing to the small number of sentences in the SQuAD documents (Wikipedia paragraphs, with five sentences on average). For the
### Table 6.10: Experimental results on TriviaQA, RACE, LearningQ and MCTest. The top three results in each metric are in bold.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>TriviaQA</th>
<th>RACE</th>
<th>LearningQ</th>
<th>MCTest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Blue 1</td>
<td>Blue 2</td>
<td>Blue 3</td>
<td>Blue 4</td>
</tr>
<tr>
<td>Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>6.69</td>
<td>2.07</td>
<td>0.70</td>
<td>0.31</td>
</tr>
<tr>
<td>Beginning</td>
<td>9.42</td>
<td>3.67</td>
<td>1.51</td>
<td>0.74</td>
</tr>
<tr>
<td>Longest</td>
<td>3.37</td>
<td>1.21</td>
<td>0.45</td>
<td>0.21</td>
</tr>
<tr>
<td>Hardest</td>
<td>1.99</td>
<td>0.66</td>
<td>0.23</td>
<td>0.11</td>
</tr>
<tr>
<td>Concept</td>
<td>0.73</td>
<td>0.19</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>ConceptType</td>
<td>1.94</td>
<td>0.57</td>
<td>0.19</td>
<td>0.08</td>
</tr>
<tr>
<td>LexRank</td>
<td>8.79</td>
<td>3.11</td>
<td>1.14</td>
<td>0.52</td>
</tr>
<tr>
<td>MMR</td>
<td>7.13</td>
<td>2.44</td>
<td>0.92</td>
<td>0.42</td>
</tr>
<tr>
<td>Novel</td>
<td>3.25</td>
<td>1.12</td>
<td>0.42</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>5.66</td>
<td>1.48</td>
<td>0.43</td>
<td>0.14</td>
</tr>
<tr>
<td>Beginning</td>
<td>5.02</td>
<td>1.29</td>
<td>0.37</td>
<td>0.13</td>
</tr>
<tr>
<td>Longest</td>
<td>6.34</td>
<td>1.81</td>
<td>0.57</td>
<td>0.22</td>
</tr>
<tr>
<td>Hardest</td>
<td>5.92</td>
<td>1.60</td>
<td>0.52</td>
<td>0.21</td>
</tr>
<tr>
<td>Concept</td>
<td>4.57</td>
<td>1.25</td>
<td>0.40</td>
<td>0.16</td>
</tr>
<tr>
<td>ConceptType</td>
<td>4.75</td>
<td>1.29</td>
<td>0.41</td>
<td>0.16</td>
</tr>
<tr>
<td>LexRank</td>
<td>6.74</td>
<td>1.91</td>
<td>0.62</td>
<td>0.26</td>
</tr>
<tr>
<td>MMR</td>
<td>5.86</td>
<td>1.53</td>
<td>0.47</td>
<td>0.17</td>
</tr>
<tr>
<td>Novel</td>
<td>6.00</td>
<td>1.64</td>
<td>0.50</td>
<td>0.17</td>
</tr>
</tbody>
</table>
other datasets, we highlight the top-3 strategies for each dataset. Based on these results, several interesting findings are observed as follows.

For the TriviaQA dataset, **Beginning** achieves the best performance, indicating that most questions in TriviaQA are about the first sentence in the source document. Considering that the articles of TriviaQA are collected from Wikipedia, such a result can be interpreted by the fact that the first sentences of Wikipedia paragraphs/articles often contain the most important information worth asking about [171]. This observation can be further verified by the well-performing results given by **LexRank** and **MRR** – ranking at the 2nd and 3rd position, respectively – which also identifies important sentences but uses a different method. Overall, these results show that importance based strategies are more effective than informativeness based (e.g., **Longest**, **Concept**), difficulty based (i.e., **Hardest**), or novelty based ones (i.e., **Novel**).

For the two datasets collected in learning contexts, namely RACE and LearningQ, **Longest**, **LexRank**, and **Novel** generally show better performance than the other strategies. Such a result suggests that questions in learning related datasets are relevant to a more diverse set of sentences, i.e., those informative, important, or contain novel information, a result we believe is due to the diverse learning goals related to the questions. We further observe big gaps between these three strategies and the remaining ones. For example, **Longest**, **LexRank**, and **Novel** are the only strategies achieving Blue 1 scores greater than 5 and Meteor scores greater than six on RACE. This observation reveals that sentence selection strategies based on similar sentence properties however measured through different textual features (e.g., **Longest** vs. **Concept** and **LexRank** vs. **Beginning**) can have big variance in terms of performance. This highlights the importance of selecting appropriate textual features in question-worthy sentence selection.

Similar results also hold on the MCTest dataset: **Longest**, **LexRank**, and **Novel** generally achieve good performance, which suggests that questions in MCTest are also relevant to a diverse set of sentences. On the other hand, strategies such as **Beginning** and **ConceptMax** also perform well on several metrics, signifying that different measures of sentence properties (e.g., informativeness using **Longest** and **ConceptMax**) do not necessarily lead to highly different sentence selection results on MCTest. Despite this, we can observe that **LexRank** is the only sentence selection strategy consistently ranking in top-3 across all the five considered datasets, demonstrating its superior robustness against all the other compared strategies.
6.7 Conclusion

In this chapter, we present LearningQ, a large-scale dataset for automatically generating educational questions by applying the state-of-the-art deep neural network approaches. It consists of 230K document-question pairs produced by both instructors and learners. To our knowledge, LearningQ is the first dataset that covers a wide range of educational topics, and the questions require a full spectrum of cognitive skills to solve. Extensive evaluation of state-of-the-art question generation methods on LearningQ shows that LearningQ is a challenging dataset that deserves significant future investigation. Moreover, we propose nine sentence selection strategies inspired by different question-asking heuristics and experiments on multiple datasets show that the beginning sentence is often worth questioning about for Wikipedia articles, while questions in learning contexts feature source sentences that are informative, important, or contain novel information.

As an implication for future research on question generation, deep neural network based methods can be further enhanced by considering the relationships among multiple source sentences and combining different strategies for selecting question-worthy sentences in question generation.
Chapter 7

Conclusion

MOOCs have been recognized as an important tool to achieve inclusive and equitable quality education and promote lifelong learning opportunities for people all over the world [121]. Typically, there are two types of MOOC platforms: *topic-agnostic* MOOC platforms like edX and Coursera provide courses covering various topics, while *topic-specific* MOOC platforms like Duolingo and Codeacademy focus on courses in one single topic. Existing research on MOOCs mainly used learner traces (e.g., video clicks, quiz submissions, forum entries) generated *within* the *topic-agnostic* MOOC platforms to investigate MOOC learning [37, 68, 72, 149]. In this thesis, we focused on (i) learner modeling and (ii) generation of educational material for both of the *topic-agnostic* and *topic-specific* MOOC platforms. In this chapter, we summarize the main contributions made in this thesis and provide an outlook on future research directions.
Chapter 7. Conclusion

7.1 Summary of Contributions

To employ the Social Web to model learners from the topic-agnostic MOOC platforms, we investigated whether MOOC learners are active in Social Web platforms and how to reliably identify these learners across multiple platforms. Concretely, in Chapter 2 we answered the following research questions:

RQ 1.1 On what Social Web platforms can a significant fraction of MOOC learners be identified?

RQ 1.2 Are learners who demonstrate specific sets of traits on the Social Web drawn to certain types of MOOCs?

RQ 1.3 To what extent do Social Web platforms enable us to observe (specific) user attributes that are highly relevant to the online learning experience?

To answer those questions, we investigated to what extent learners from eighteen MOOCs in edX could be discovered across five popular Social Web platforms (i.e., Gravatar, Twitter, LinkedIn, StackExchange and GitHub) and further derived a set of learner attributes from these platforms to investigate learners’ behaviors in MOOCs. Depending on the MOOC-platform combination, we identified between 1% and 42% of learners (5% on average) in the five considered platforms (RQ 1.1). In the most extreme case, 42% of learners from a Functional Programming MOOC could be identified in GitHub. We also showed that learners with specific traits were attracted to different types of MOOCs (RQ 1.2). In particular, we presented a first investigation into the knowledge application behavior of learners, i.e., learning transfer, beyond the MOOC platform over time (RQ 1.3). We provided a reliable methodology to gather information about learners by moving from the MOOC platform to the wider Social Web. More importantly, we demonstrated that a set of valuable learner attributes relevant to MOOC learning can be derived from the Social Web. The data-driven approaches used in our work can be applied in not only the MOOC setting but also other educational settings like e-learning courses as well as campus-based courses, as long as the learners can be identified in the Social Web.

After observing that learners of programming courses actively engaged with GitHub, we considered the Functional Programming MOOC as a specific case and continued the investigation of learning transfer. Concretely, in Chapter 3 we answered the following research questions:
RQ 2.1 To what extent do learners from a programming MOOC transfer the newly gained knowledge to practice?

RQ 2.2 What type of learners are most likely to make the transfer?

RQ 2.3 How does the transfer manifest itself over time?

To answer those questions, we conducted a large-scale longitudinal analysis, in which both the learning traces generated within the MOOC platform and the coding traces collected from GitHub were used. We observed that about 8% of engaged learners, who had no prior knowledge in functional programming, began programming functionally after the MOOC (RQ 2.1). In addition, learners were more likely to make the transfer if they had (i) intrinsic motivation, (ii) high self-efficacy, (iii) prior experience in programming, and (iv) a high personal capacity (RQ 2.2). Lastly, neither a significant transfer increase nor decrease was observed over half a year after the course (RQ 2.3). By examining programming learners’ uptake of knowledge after the MOOC, instructors can not only gain a better understanding of the course influence on learners but also evaluate the current course and design future courses to induce more knowledge transfer.

Most existing research focused on investigating learner behaviors in topic-agnostic MOOC platforms. We used the three large-scale language learning datasets, which were released by Duolingo in the Second Language Acquisition Modeling challenge, to enable a better understanding of learners in the topic-specific MOOC platforms. Concretely, in Chapter 4 we answered the following research question:

RQ 3.1 What factors are correlated with learners’ language learning performance?

To answer the question, we conducted an analysis on the three Duolingo datasets and demonstrated that factors like the amount of time spent in learning and the devices being used were related to learners’ accuracy in solving exercises and the amount of vocabulary learned. Furthermore, based on the results, we designed a set of features and examined their effectiveness in predicting learners’ future performance in the setting of second language acquisition.

As demonstrated in Chapter 3, learners indeed transferred the knowledge acquired from a MOOC to practice. We further investigated whether learners could apply the acquired knowledge to solve real-world tasks, i.e., paid tasks
collected from online marketplaces which can be solved by learning with the course. If learners are capable of solving such tasks, ultimately, we envision to construct a recommender system to suggest learners solve paid freelancing tasks relevant to the course, as a possible means to earn money when learning with the course. Concretely, in Chapter 5 we answered the following research questions:

**RQ 4.1** Are MOOC learners able to solve real-world (paid) tasks from an online work platform with sufficient accuracy and quality?

**RQ 4.2** How applicable is the knowledge gained from MOOCs to paid tasks offered by online work platforms?

**RQ 4.3** To what extent can an online work platform support MOOC learners (i.e., are there enough tasks available for everyone)?

**RQ 4.4** What role do real-world (paid) tasks play in the engagement of MOOC learners?

To answer those questions, we designed a study, in which we manually selected a set of tasks from Upwork and deployed them into a MOOC teaching data analysis as bonus exercises for learners to solve. We demonstrated that learners could solve the paid tasks with the knowledge gained from the course with high accuracy and quality (**RQ 4.1 & RQ 4.2**). However, there were not sufficient tasks available on Upwork to sustain the learner population throughout the entire run of the course (**RQ 4.3**). We also observed that real-world tasks were likely to have a positive effect on learners’ course engagement (**RQ 4.4**). Our study contributed the first step to develop a paid task recommender systems that we envisioned to help learners earn money when learning with a MOOC. With more online marketplace platforms considered and a larger number of paid tasks retrieved, we hypothesize the proposed system can truly help learners, especially learners who suffer from poor financial situations and consequently have a limited amount of time for learning because of the need to work and earn a living, to gain more time to learn with MOOCs.

Driven by the importance of questions in learning and the need of easing instructors’ burden in manually creating a large question bank to meet the needs of various learners, we explored the Social Web to collect a large-scale educational question dataset. With the collected dataset, we investigated whether an educational question generator could be constructed and how to effectively select question-worthy sentences from an article. Concretely, in Chapter 6 we answered the following research questions:
7.2. Future Work

**RQ 5.1** Can a large-scale and high-quality educational question dataset be collected from the Social Web?

**RQ 5.2** What are effective strategies in identifying question-worthy sentences from an article?

To answer those questions, we turned to education-oriented Social Web platforms. Specifically, we targeted TED-Ed and Khan Academy as our main data sources and collected a large-scale educational question dataset (LearningQ), which consists of over 230K document-question pairs generated by both instructors and learners (RQ 5.1). In particular, the questions contained in LearningQ vary in all cognitive levels in the Bloom’s Revised Taxonomy and cover a wide range of learning topics. With LearningQ as a testbed, we demonstrated the research challenges in constructing an educational question generator and examined the effectiveness of nine strategies in selecting question-worthy sentences from an article for educational question generation (RQ 5.2).

7.2 Future Work

This thesis has contributed novel technical approaches to model learners and generate educational material for both topic-agnostic and topic-specific MOOC platforms. However, there is still space for improvements. In this section, we provide an outlook on interesting research directions in MOOCs opened up by the research conducted in this thesis.

7.2.1 Adaptive Learning in MOOCs

Adaptive learning is an educational approach which employs computational algorithms to decide what learning materials should be presented to a learner so as to address the unique needs of the learner [3, 110, 120, 132]. Though being recognized as essential by instructors, adaptive learning has not been fully supported and investigated in MOOCs yet. One key step before enabling effective adaptive learning is to construct learner models, which are built based on learner data. However, there is no data about learners available in the learning platform at the beginning of a course unless they have been active in the platform before and, even then, the learners’ knowledge on the course topic might still be unknown to the instructors.
Our works presented in Chapters 2-3 have demonstrated that it is possible to enhance the construction of a learner model by mining the Social Web. To further the learner model construction in MOOCs, in addition to the eight Social Web platforms investigated in this thesis, it will be valuable to explore other Social Web platforms (e.g., YouTube, Instagram, Quora) to reveal a more diverse set of learner attributes in the future. These attributes could be learners’ interests, prior knowledge, learning preferences, personal goals, social relations, and so on. Building upon the enhanced learner models, future work can also focus on developing effective adaptive algorithms to personalize learner experiences in MOOCs. For instance, what learner attributes should be considered to generate a personalized learning path for a MOOC learner? What are the influences of prior knowledge on learners’ learning paths? How can personalized learning strategies (e.g., tips in time management) be generated based on the learner models? How can the temporal dynamics of learner behaviors be captured and used to provide adaptive learning supports (e.g., recommended learning paths, personalized learning strategies)?

7.2.2 Interactive Learning in MOOCs

Interactive learning is a pedagogical model which encourages learners to interact with each other instead of passively absorbing the knowledge taught in the course. In the classroom setting, interactive learning occurs in a variety of forms such as hands-on group projects and class discussions. However, in the MOOC setting, learning takes place in an asynchronous manner, and learners’ interaction with instructors and peers are mostly limited to the discussion forum. Thus, learners cannot gain a wealth of experience in interactive learning.

Future research on developing interactive tools for MOOC learners can be built on the works presented in this thesis. Specifically, with the data collected from the Social Web, e.g., the LearningQ in Chapter 6, an intelligent personal assistant can be constructed and used to help MOOC learners by providing the support they need. The support could be: discussing questions, scheduling time for learning, providing emotional support, and so on. Future research can first work on investigating what learning support is needed in the MOOC setting. For each kind of support, it will be valuable to explore what data and techniques can be used to enable the support. Furthermore, how should the interface of the assistant be designed so as to engage learners? What are effective strategies to allow the assistant to interact with learners?
7.2. Future Work

7.2.3 Content Enrichment in MOOCs

Most existing MOOCs in the major topic-agnostic platforms adopt the one-size-fits-all approach, i.e., providing the same set of learning materials to all learners in a MOOC. However, these learners are often of high diversity (e.g., their demographics [30, 74, 177]) and likely to have different learning needs, which demand the MOOC to contain a larger and more diverse set of learning materials instead of only a limited number of videos and quizzes to meet their needs. To enrich MOOC content, there have been several studies working on Learnersourcing [61], i.e., employing the intelligence of learners enrolled in a MOOC to gather or create more content for the MOOC. However, Learnersourcing faces the problem of lacking enough responses from learners as such content gathering and creation process is very cognitively demanding and time-consuming.

Our works presented in Chapters 5-6 demonstrated the potential of the Social Web in enriching MOOCs. Building on the work presented in Chapter 5, future research can focus on developing techniques to automatically retrieve relevant freelancing tasks and determine their relevance and suitability to a course. Considering that a large number of freelancing tasks are with high payment and more challenging to solve, it will be interesting to investigate how to enable the partition of a high-payment task across several learners. Ideally, such partition can motivate the learners to learn and solve the task together and then the learners can share the payment. In Chapter 6, we focused on generating text-based questions with the aid of LearningQ. In addition to text-based questions, future research can also work on the generation of questions consisting of not only text but also plots and images, which are necessary for course topics like math and physics. Furthermore, how can the answers provided by learners to the generated questions be automatically assessed? As demonstrated in Chapter 6, educational content creation is a cognitively demanding task, which we believe cannot be achieved by simply exploiting the power of machines in the near future. We hypothesize that MOOC content enrichment can be greatly enhanced by combining the power of humans and machines, e.g., using human intelligence as a means to refine low-quality questions generated by the algorithms. How can learners effectively assist machines to create course content? What are the effects on learning by enabling learners to create course content? How can the usefulness of the created content be measured? How can the feedback provided by learners in the process of content creation be effectively utilized (e.g., by applying reinforcement learning [145]) to further improve the performance of algorithms?
Bibliography


[44] Peter de Vries and Thieme Hennis. Tu delft online learning research working paper #6, 2014.


[73] Thomas Hill, Nancy D. Smith, and Millard F. Mann. Role of efficacy expectations in predicting the decision to use advanced technologies:


[94] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. The knowledge-learning-instruction framework: Bridging the science-


List of Figures

1.1 An overview of the MOOC stages and data sources investigated in Chapters 2-6. The number in a cell represents the corresponding chapter, which focuses on the MOOC stage specified in the column and the MOOC platform or the Social Web platform specified in the row ............................................. 3

2.1 Excerpt of a GitHub PushEvent log trace ................................. 18

2.2 Percentage of our Twitter users across eight age brackets. The “Overall” user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific .................................................. 23

2.3 Percentage of our Twitter users of each gender. The “Overall” user set contains all users independent of the specific MOOC(s) taken, the remaining three user sets are MOOC-specific ........ 23

2.4 Overview of the most frequent job title bigrams among the learners of the Data Analysis (top), Delft Design Approach (middle), and Responsible Innovation (bottom) MOOCs .... 24

2.5 Fraction of learners displaying n numbers of MOOC certificate 25

2.6 t-SNE based visualization of LinkedIn skill vectors for pairs of MOOCs. Each data point represents one skill vector (i.e. one user). .......................................................... 27
2.7 Overview of the number of StackOverflow questions and answers posted on a monthly basis between January 2014 and July 2015 by (i) our MOOC learners [top and middle], and (ii) all StackExchange users [bottom] for Haskell [top, bottom] and the nine major functional languages [middle]. Marked in gray is the time period of the Functional Programming MOOC. The dashed green line indicates the ratio of Questions:Answers in each month. ................................. 29

2.8 Month-by-month GitHub contributions in the Haskell language by the Functional Programming MOOC learners identified on GitHub. ........................................ 30

3.1 Excerpt of a GitHub PushEvent log trace .......................... 43

3.2 Excerpt of a diff file. Two files were changed (viewsA.rb and routes.rb). The extension * .rb indicates code written in Ruby. 44

3.3 Number of unique users actively using a functional language. FP101x ran during the highlighted region .............................. 48

3.4 Fraction of functional programming activities among the 1,163 engaged Expert Learners. FP101x ran during the highlighted region. ........................................ 49

3.5 Fraction of functional programming activities among the 542 engaged Expert Learners with functional activities before & after FP101x. FP101x ran during the highlighted region. . . . 50

3.6 Fraction of functional programming activities among the 336 engaged Novice Learners with functional activities after FP101x FP101x ran during the highlighted region. .......................... 51

3.7 Distribution of functional programming activities among the 336 engaged Novice Learners with functional activities after FP101x. FP101x ran during the highlighted region. .......................... 52

3.8 Functional languages used by the 336 engaged Novice Learners during and after FP101x. Best viewed in color .......................... 53

3.9 Fraction of the 336 engaged Novice learners remaining active in each course week. 70 Novice learners completed FP101x successfully, 266 did not complete it. ................................. 54
4.1 The average learner-level accuracy, i.e., Lear-Acc (Top), and the average number of mastered words, i.e., Mast-Word (Bottom), of learners in high-spacing and low-spacing groups

5.1 Paying MOOC learners — a vision

5.2 Bonus exercise posted in week 4 of EX101x. The original task was posted with a price of $35 to Upwork (note that at the time of posting this exercise, Upwork was still called oDesk)

5.3 Developed countries according to the OECD are shown in blue, developing countries are shown in red. The color shade indicates the overall completion rate of learners from that country. A darker shade indicates a higher completion rate

5.4 From the 56,308 Upwork tasks available on 15/09/2015 a total of 8,153 have a fixed budget (the remaining tasks are paid by the hour). Budgeted tasks are binned according to the budget they have

5.5 The 56,308 Upwork tasks available on 15/09/2015 are binned according to the number of days they have been “online” (i.e., the task is open)

5.6 The average amount of time (in hours) that learners spent in watching video after viewing (but not submitting) the bonus exercises. The numbers of learners within each group are given in brackets. Results marked with * $(p < 0.001)$ are significantly different (Viewed vs. Not viewed) according to the Mann-Whitney U-test

5.7 Overview of the freelance work task recommender system’s design

6.1 Results of question generation on different learning subjects in Khan Academy

6.2 Results of question generation with different source sentence lengths
List of Tables

2.1 Overview of the edX MOOCs under investigation, the number of learners registered to those MOOCs and the number of learners that could be matched (with either Explicit/Direct or Fuzzy matching) to our five Social Web platforms. Marked with ↑ (†) are those course/platform combinations where we were able to locate > 5% (> 10%) of the registered learners. The final row contains the unique number of users/learners (a learner may have taken several MOOCs) identified on each platform. .......................................................... 20

2.2 Overview of the percentage of MOOC learners (329,200 overall) identified through the different matching strategies on the five selected Social Web platforms. A dash (—) indicates that for this specific platform/strategy combination, no matching was performed. .......................................................... 21

3.1 Overview of the different data sources used to investigate each research hypothesis. CS refers to the conducted Course Surveys (before and after the course). ........................................... 40

3.2 Basic characteristics across all learners and their partitioning into GitHub (GH) and non-GitHub learners. Significant differences (according to Mann-Whitney) between GH and non-GH learners are marked with †(p < 0.001). .................................................. 46

3.3 Basic characteristics when partitioning the GitHub learners according to prior functional programming expertise. Significant differences (according to Mann-Whitney) between Expert and Non-Expert learners are marked with †(p < 0.001) and ‡(p < 0.01). .................................................. 47
3.4 Partitioning of the 336 Novice learners according to several
dimensions. The last column shows the number of learners
that could not be assigned (N/A) to a dimension. 51

3.5 The number of Novice Learners falling into spacing groups. 52

4.1 Statistics of the datasets. 59

4.2 Avg. learner-level accuracy (%) and the number of mastered
words of learners living in different locations (approximated by
the countries from which learners have finished the exercises).
Significant differences (compared to Avg., according to Mann-
Whitney) are marked with * (p < 0.001). 62

4.3 Pearson Correlation between learner engagement (measured
by # attempted exercises/words and the amount of time spent
in learning) and learner-level accuracy as well as # mastered
words. Significant differences are marked with * (p < 0.001). 62

4.4 Pearson Correlation between the amount of time spent in solv-
ing each exercise and exercise-level accuracy. Significant dif-
fferences are marked with * (p < 0.001). 63

4.5 Average exercise-level accuracy (%) in different contextual con-
ditions. Significant differences (compared to Avg., according
to Mann-Whitney) are marked with *(p < 0.001). 64

4.6 Avg. word-level accuracy (%) of words with different number
of exposures. 65

4.7 Pearson Correlation between learner performance and the num-
ber of previous attempts and the amount of time elapsed since
the last attempt for a word. 65

4.8 Granularity levels on which each feature is retrieved or com-
puted. Features marked with b are used as input in the baseline
provided by the benchmark organizers. 68

4.9 Model parameters of the GTB model; determined by using
grid search per dataset. 71

4.10 Experimental results reported in AUC on ES-EN. Each row
indicates a feature added to the GTB feature space; the model
of row 1 has three features. 72
5.1 Basic characteristics across all learners and their partitioning into those who attempted to solve at least one Bonus Exercise (BE) and those who did not (Non-BE). Where suitable, significance tests between the BE/Non-BE groups were performed according to Mann-Whitney. All performed tests exhibited significant differences - indicated with \( \dagger \) (significant difference with \( p < 0.001 \)).

5.2 Basic characteristics of BE learners partitioned into dedicated BE learners (DBE) solving 3+ bonus exercises and non-dedicated BE learners. Where suitable, significance tests between the DBE/Non-DBE groups were performed according to Mann-Whitney. All performed tests exhibited significant differences - indicated with \( \ddagger \) (significant difference with \( p < 0.001 \)).

5.3 Learners' performance on real-world tasks. The second column shows the number of active learners. The third column shows the number of students taking the bonus exercise. The fourth column shows the task payment offered at UpWork. Accurate submissions are those matching our gold standard (with the additional requirement of the correct order for tasks 3 and 5). High-quality submissions are those correct submissions without code smells. The coverage column reports the average (and standard deviation) fraction of cells covered by all of a week's submissions.

5.4 Paid total worker fees by company in Million US Dollar. These numbers are self reported by the companies and are not given for a specific year.

5.5 Overview of programming tasks among our crawl of 56,308 Upwork tasks on 15/09/2015.

5.6 The 56,308 Upwork tasks available on 15/09/2015 are partitioned according to their category. Shown are the number of tasks per category, the average number of days online and the average task payment (for the subset of 8,153 tasks with a fixed budget).
5.7 Overview of the 11 questions in our post-course survey. For presentation purposes, some questions and answers appear slightly condensed. For all closed-form questions, we provide the distribution of answers (in %) across the four learner partitions in the form A | B | C | D%: (A) from developed nations + at least one bonus exercise submitted, (B) from developing nations + at least one bonus exercise submitted, (C) from developed nations + no bonus exercise submitted, and, (D) from developing nations + no bonus exercise submitted.

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Examples of document-question pairs</td>
</tr>
<tr>
<td>6.2</td>
<td>Question-worthy sentence in a paragraph</td>
</tr>
<tr>
<td>6.3</td>
<td>Examples of useful (marked with ✓) and non-useful questions from Khan Academy. S/H/M/C/E/T denote Science, Humanities, Math, Computing, Economics and Test Preparation, respectively.</td>
</tr>
<tr>
<td>6.4</td>
<td>Descriptive features and statistics of LearningQ and the datasets in comparison</td>
</tr>
<tr>
<td>6.5</td>
<td>Top words in documents and questions and top interrogative words of questions in LearningQ and the datasets in comparison. Words pertinent to a specific data source platforms are in bold. KA represents Khan Academy</td>
</tr>
<tr>
<td>6.6</td>
<td>Question Examples of Different Bloom’s Revised Taxonomy Level in TED-Ed and Khan Academy</td>
</tr>
<tr>
<td>6.7</td>
<td>Distribution of Bloom’s Revised Taxonomy Labels</td>
</tr>
<tr>
<td>6.8</td>
<td>Results of Source Sentence Labelling. # Words/Sent. denote the average words/sentences in the labelled source sentences. % ONE/MULTIPLE/EXTERNAL refer to the percentage of questions related to ONE single sentence, MULTIPLE sentences or require EXTERNAL knowledge to generate, respectively. KA denotes Khan Academy</td>
</tr>
<tr>
<td>6.9</td>
<td>Performance of rule-based and deep neural network based methods on LearningQ</td>
</tr>
<tr>
<td>6.10</td>
<td>Experimental results on TriviaQA, RACE, LearningQ and MCTest. The top three results in each metric are in bold</td>
</tr>
</tbody>
</table>
Summary

MOOC Analytics: Learner Modeling and Content Generation

Massive Open Online Courses (MOOCs), as one of the popular options for people to receive education and learn, are endowed with the mission to educate the world. Typically, there are two types of MOOC platforms: topic-agnostic and topic-specific. Topic-agnostic platforms such as edX and Coursera provide courses covering a wide range of topics, while topic-specific MOOC platforms such as Duolingo and Codeacademy focus on courses in one specific topic. To better support MOOC learners, many works have been proposed to investigate MOOC learning in the past decade. Still, there are many other aspects of MOOC learning to be explored.

In this thesis, we focused on (i) learner modeling and (ii) generation of educational material for both topic-agnostic and topic-specific MOOC platforms.

For learner modeling in the topic-agnostic platforms, as there have been a lot of works utilizing the learner traces generated within the MOOC platforms, we proposed that we can better understand learners by moving beyond the MOOC platforms and exploring other data sources on the wider Web, especially the Social Web. As an exploratory but necessary step, in Chapter 2, we first investigated whether MOOC learners are active in the Social Web and how to reliably identify their accounts across various Social Web platforms. To this end, we considered over 320,000 learners from eighteen MOOCs in edX and made efforts to identify their accounts across five popular Social Web platforms, i.e., Gravatar, Twitter, LinkedIn, StackExchange and GitHub. Furthermore, we investigated what data traces could be collected from these platforms and used to derive learner attributes that are relevant to their learning activities in the MOOC setting. We found that on av-
erage 5% of learners could be identified on globally popular Social Web platforms and learners with specific traits preferred different types of MOOCs. Based on the observations we had in Chapter 2, in which we have observed that over one-third of learners from a Functional Programming MOOC used GitHub to maintain their programming activities, we further combined the data traces generated by those learners in both edX and GitHub to investigate learning transfer in Chapter 3. Do learners actually use the newly acquired knowledge and skills to solve problems in practice? Our analyses revealed that (i) more than 8% of engaged learners transferred the acquired knowledge to practice, and (ii) most existing transfer learning findings from the classroom setting are indeed applicable in the MOOC setting as well. For learner modeling in the topic-specific platforms, in Chapter 4, we focused on investigating the problem of knowledge tracing, which remained largely unexplored in previous studies due to the lack of available datasets from such platforms. With three large-scale language learning datasets released by Duolingo, we investigated factors that are correlated with learners’ performance and then applied a machine learning technique (i.e., Gradient Tree Boosting) to predict learners’ future performance. We demonstrated that the learning performance was correlated with not only learners’ engagement with a course but also contextual factors like the devices being used. In Chapter 5, we further investigated whether learners could apply the acquired knowledge to solve real-world tasks, i.e., paid tasks which are retrieved from online marketplaces and can be solved by applying the knowledge taught in a course. For this purpose, we considered a MOOC teaching data analysis in edX and manually selected a set of paid tasks from Upwork, one of the most popular freelancing marketplaces in the world, and presented the selected tasks to learners and observed how learners interacted with these real-world tasks. We observed that these tasks could be solved by MOOC learners with high accuracy and quality.

For generation of educational material, in Chapter 6, we focused on the generation of educational questions, as they are widely recognized as essential for learning. To build an effective automatic question generator, two challenges need to be overcome. Firstly, a large-scale dataset covering questions of various cognitive levels from a set of diverse learning topics should be collected. Secondly, effective strategies for identifying question-worthy sentences (i.e., those carrying important concepts) from an article, should be developed before using those sentences as input to the question generator. To deal with these challenges, we relied on TED-Ed and Khan Academy to retrieve an educational question dataset, LearningQ, which contains over 230K document-question pairs generated by both instructors and learners. We
showed that *LearningQ* consists of high-quality questions covering not only all cognitive levels in the Bloom’s Revised Taxonomy but also various learning topics. We showed that it is a challenging task to automatically generate educational questions, even with sufficient training data and state-of-the-art question generation techniques. Besides, we developed and compared a total of nine strategies to select question-worthy sentences from an article and demonstrated that questions in learning contexts usually are based on source sentences that are informative, important, or contain novel information.
Samenvatting

MOOC Analyse: Modelleren van studenten en genereren van content

Massive Open Online Courses (MOOC’s) zijn, als een van de populaire manieren waarop mensen onderwijs krijgen en leren, verbonden met de missie om de wereld te onderwijzen. Karakteristiek zijn er twee typen van MOOC-platforms: onderwerp-onafhankelijk en onderwerp-specifiek. Onderwerp-onafhankelijke platforms zoals edX en Coursera bieden cursussen aan over een breed spectrum van onderwerpen, terwijl onderwerp-specifieke MOOC-platforms zoals Duolingo en Codeacademy zich richten op cursussen in een specifiek onderwerp. Om MOOC-studenten beter te ondersteunen is er in het afgelopen decennium veel onderzoek gedaan naar het leren in MOOC’s. Desondanks zijn er nog veel aspecten van het leren in MOOC’s die nog moeten worden onderzocht.

In dit proefschrift richten we ons op (i) modelleren van studenten en (ii) genereren van educatieve content voor zowel onderwerp-onafhankelijke als onderwerp-specifieke MOOC-platforms.

Voor het modelleren van studenten in de onderwerp-onafhankelijke platforms, hebben we voorgesteld dat we studenten beter kunnen begrijpen door verder te kijken dan de MOOC-platforms en andere databronnen op het wijdere Web te verkennen, speciaal het Social Web; dit omdat een heleboel onderzoek al de student-logs hebben benut die door de MOOC-platforms zelf worden gegenereerd. Als een verkennende maar noodzakelijke stap, hebben we in hoofdstuk 2 eerst onderzocht of MOOC-studenten actief zijn in het Social Web en hoe we betrouwbaar hun accounts op verschillende Social Web platforms kunnen identificeren. Hiertoe hebben we meer dan 320.000 studenten van achttien MOOC’s in edX onderzocht en gekeken hoe we hun accounts op vijf populaire Social Web platforms kunnen identificeren, i.c.,
Samenvatting

Gravatar, Twitter, LinkedIn, StackExchange en GitHub. Verder hebben we onderzocht welke logs van deze platforms kunnen worden verzameld en benut om student-attributen af te leiden die relevant zijn voor hun leeractiviteiten in de MOOC-context. We stelden vast dat gemiddeld 5% van de studenten konden worden geïdentificeerd op globaal populaire Social Web platforms en dat studenten met specifieke eigenschappen voorkeur hebben voor verschillende typen van MOOC’s. Gebaseerd op de observaties van Hoofdstuk 2, waar we observeerden dat meer dan een derde van de studenten van een Functional Programming MOOC GitHub gebruikten voor hun programmeer-activiteiten, combineerden we de logs gegenereerd door die studenten in zowel edX als GitHub om zogenoemde learning transfer te onderzoeken in Hoofdstuk 3: gebruiken studenten daadwerkelijk de nieuw verworven kennis en vaardigheden voor problemen in de praktijk? Onze analyses toonden aan dat (i) meer dan 8% van betrokken studenten inderdaad de verworven kennis benutten in de praktijk, en (ii) de meeste bestaande inzichten over transfer van de klassieke klas-context inderdaad ook van toepassing zijn in de MOOC-context. Voor het modelleren van studenten in onderwerp-specifieke platforms, hebben we ons in Hoofdstuk 4 gericht op het onderzoeken van het probleem van kennis-herleiding, dat grotendeels niet is onderzocht in eerdere onderzoeken vanwege het gebrek aan beschikbare datasets van zulke platforms. Met drie grootschalige datasets rond taalverwerving beschikbaar gesteld door Duolingo hebben we factoren onderzocht die gecorreleerd zijn met de performance van studenten en dan een machine learning-techniek toegepast (namelijk Gradient Tree Boosting) om de toekomstige performance van studenten te voorspellen. We hebben aangetoond dat de performance van de studenten was gecorreleerd met niet alleen de betrokkenheid van de student met de cursus maar ook met contextuele factoren zoals de apparaten die werden gebruikt. In Hoofdstuk 5 hebben we verder onderzocht of studenten de verkregen kennis zouden kunnen toepassen in echte realistische taken, i.e. betaalde taken verkregen van online marktplaatsen die kunnen worden volbracht door de kennis toe te passen uit de cursus. Voor dit doel hebben we een data-analyse van MOOC-onderwijs in edX beschouwd en handmatig een set van betaalde taken geselecteerd van Upwork, een van de populairste freelance marktplaatsen ter wereld, en de geselecteerde taken aan studenten aangeboden en gezien hoe de studenten met deze echte realistische taken omgaan. We hebben geconstateerd dat deze taken konden worden volbracht door MOOC-studenten met hoge accuratesse en kwaliteit.

Voor het genereren van educatieve content hebben we ons in Hoofdstuk 6 gericht op het genereren van educatieve vragen, aangezien die algemeen als essentieel worden beschouwd voor leren. Om een effectieve automatische
Samenvatting

vraag-generator te bouwen, moeten twee uitdagingen worden overwonnen. Ten eerste moet er een grootschalige dataset worden verzameld van vragen van verschillende kennisniveaus voor verschillende onderwerpen. Ten tweede moeten effectieve strategieën worden ontwikkeld voor het identificeren van zinnen in een artikel die een vraag waard zijn (i.e. zinnen die belangrijke concepten bevatten), voordat deze zinnen worden gebruikt als input voor een vraag-generator. Om met deze uitdagingen om te gaan, hebben we ons gebaseerd op TED-Ed en Khan Academy om een educatieve dataset van vragen te verkrijgen, LearningQ, met meer dan 230K document-vraag-paren gegenereerd door zowel docenten als studenten. We hebben aangetoond dat LearningQ bestaat uit vragen van hoge kwaliteit die niet alleen alle kennisniveaus van Bloom’s Revised Taxonomy beslaan maar ook verschillende onderwerpen. We hebben laten zien dat het een uitdagende taak is om automatisch educatieve vragen te genereren, zelfs met voldoende trainingsdata en state-of-the-art technieken voor vraag-generatie. Daarnaast hebben we in totaal negen strategieën ontwikkeld en vergeleken om uit een artikel zinnen te selecteren die een vraag waard zijn en aangetoond dat vragen in leercontexten doorgaans gebaseerd zijn op bronzinnen die informatief zijn, belangrijk zijn en nieuwe informatie bevatten.
Curriculum Vitae

Guanliang Chen was born in Zhanjiang, China on February 24, 1988. He received his master degree with an outstanding thesis award from South China University of Technology, China. His master thesis was completed when he served as an exchange research student at Hong Kong Baptist University, China, where he worked on developing effective context-aware recommender systems. Prior to that, he received his bachelor degree from South China University of Technology, China.

From March 2015 to December 2018, Guanliang Chen was a PhD student in the Web Information Systems group at Delft University of Technology, supervised by Geert-Jan Houben and Claudia Hauff. His PhD work focused on developing data-driven approaches for better modeling MOOC learners and generating useful educational content through external sources, in particular, the Social Web. Guanliang’s research has been published in leading conferences and journals on relevant fields such as ICWSM, WebScience, UMAP, L@S, LAK, EDM, EC-TEL, IEEE Transactions on Learning Technologies, and Computers & Education. He received Best Student Paper Award from EC-TEL 2016 and Best Student Paper Nominee Award from both L@S 2016 and UMAP 2014. Guanliang co-organized the Workshop on Integrated Learning Analytics of MOOC Post-Course Development at LAK 2017 and the 15th Dutch-Belgian Information Retrieval Workshop in 2016. He was invited as a keynote speaker in the 2017 Doctoral Student Forum on MOOC Research in Peking University, China. He has also served as a program committee member and reviewer for several conferences and journals, such as LAK, ICWL, IEEE Transactions on Learning Technologies, Computers & Education, ACM Computing Surveys, etc.
Publications


8. Dan Davis, Ioana Jivet, René F. Kizilcec, Guanliang Chen, Claudia Hauff, Geert-Jan Houben. Follow the Successful Crowd: Raising
MOOC Completion Rates through Social Comparison at Scale. In Proceedings of the 7th International Conference on Learning Analytics and Knowledge, Vancouver, Canada. LAK’17. (Full conference paper)


SIKS Dissertation Series

Since 1998, all dissertations written by Ph.D. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series.

2019-13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
2019-12 Jacqueline Heinerman (VU), Better Together
2019-11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
2019-10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
2019-09 Payameh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
2019-08 Fei de Nije (TUD), Resource-constrained Multi-agent Markov Decision Processes
2019-07 Soude Fazeli (OUN), Recommender Systems in Social Learning Platforms
2019-06 Chris Dijkshoorn (VUA), Niche searching for Improving Access to Linked Cultural Heritage Datasets
2019-05 Sebatiaan van Zelst (TUE), Process Mining with Streaming Data
2019-04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
2019-03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
2019-02 Emmanuelle Beauxis-Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
2019-01 Rob van Eijk (UL), Comparing and Aligning Process Representations
2018-30 Wouter Boek (VU), The “K” in “semantic web”: the role of “knowledge”, making semantics to the web
2018-29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech
2018-28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
2018-27 Mathi Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
2018-26 Rosolof de Vries (U), Theory-Based And Task-Based: Motivational Messages for Behavior Change Technology
2018-25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
2018-23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
2018-22 Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviors, Perceptions and Emotions in Social Networks
2018-21 Aad Schotmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
2018-20 Manxia Liu (RUN), Time and Bayesian Networks
2018-19 Minh Due Pham (VUA), Emergent relational schemas for RDF
2018-18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
2018-17 Jianpeng Zhang (TUE), On Graph Sample Chaining
2018-16 Jacobok Kim (UT), Automatic recognition of engagement and emotion in a group of children
2018-15 Naeer Dassaranaa (UM), Biomarker discovery in heart failure
2018-14 Bart Joosten (UT), Detecting Social Signals with Spatiotemporal Gabor Filters
2018-13 Seyed Amin Tabatabaei (UVA), Using behavioral context in process mining: Exploring the added value of computational models for increasing the use of renewable energy in the residential sector
2018-12 Xici Lu (TUE), Using behavioral context in process mining
2018-11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
2018-10 Julienka Mollee (VUA), Moving forward: supporting physical activity behaviour change through intelligent technology
2018-09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
2018-08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
2018-07 Jutong Luo (UU), A formal account of opportunism in multi-agent systems
2018-06 Dan Ioitia (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
2018-05 Hugg Houckelman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
2018-04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
2018-03 Steven Bosena (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
2018-02 Felix Mannhardt (TUE), Multi-perspective Process Mining
2018-01 Han van der Aa (VUA), Comparing and Aligning Process Representations
2017-48 Angel Suarez (OU), Collaborative inquiry-based learning
2017-47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
2017-46 Jan Schneider (OU), Sensor-based Learning Support
2017-45 Bas Tostering (U), Decentralized Runtime Norm Enforcement
2017-44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
2017-43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
2016-22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
2016-21 Alejandro Moreno Cleri (UT), From Traditional to Interactive Playgrounds: Automatic Analysis of Player Behavior in the Interactive Top Playground
2016-20 Daan Olthof (UVA), Context & Semantics in News F/B Search
2016-19 Julia Eitemova (TU), Mining Social Structures from Genealogical Data
2016-18 Albert Moreo-Pena (VU), Refining Statistical Data on the Web
2016-17 Berend Wold (VU), Towards Embodied Evolution of Robot Organisms
2016-16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
2016-15 Steffen Michele (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
2016-14 Ravi Khalka (UU), Revisiting Legacy Software System Modernization
2016-13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICTD Approach
2016-12 Max Knobhont (UVA), Logics for Modelling and Verifying Normative Multi-Agent Systems
2016-11 Anne Schult (UVA), Search Engines that Learn from Their Users
2016-10 George Karastattis (VUVA), Parameter Control for Evolutionary Algorithms
2016-09 Archana Vaithiyalingam (VU), Trusting Crowdsourced Information on Cultural Artifacts
2016-08 Matje van de Camp (TU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
2016-07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
2016-06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
2016-05 Evgeny Shekhorov (UVA), Expanded Argu- que: Containment and an Application in Explaining Missing Answers
2016-04 Laurens Rietveld (VU), Publishing and Connecting Linked Data
2016-03 Maria Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
2016-02 Michel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: presenting a better pill to swallow
2016-01 Syed Muslim Abbas (RUN), Recognition of Shapes by Humans and Machines
2016-00 Jungao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction
2015-34 Victor de Graaf (UT), Gesocial Recommender Systems
2015-33 Frederik Schaud (TUD), Ontology Mapping with Auxiliary Resources
2015-32 Jorn van den Elp (U), Corporate Venture Management in SMEs
2015-31 Jukka Kii (TUD), On the robustness of Power Grids
2015-30 Kiaash Bahreini (UU), Real-time Multimodal Emotion Recognition in E-Learning
2015-29 Hendrik Baire (UU), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains
2015-28 Janet Bagorogou (CU), KNOWLEDGE MANAGEMENT AND HIGH PERFORMANCE: The Uganda Financial Institutions Model for HPO
2015-27 Sándor Háman (CWI), Updating compressed volume stores
2015-26 Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure
2015-25 Steven Woudenborg (UU), Bayesian Tools for Early Disease Detection
2015-23 Luit Gaudzic (VU), Catalogue Support in Cultural Heritage: Multi-Agent Systems
2015-22 Zhebin Zhu (UT), Co-occurrence Rate Networks
2015-21 Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning
2015-20 Loes Vanhui (UU), Using Culture and Values to Support Flexible Coordination
2015-19 Bernardo Talmaciu (OUN), Ubiquitous Technology for Lifelong Learners
2015-18 Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories
2015-17 André van Cleeff (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs
2015-16 Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork
2015-15 Klaas Audretsch (TUD), Ontology-based Software Architecture Documentation
2015-14 Bart van Straalen (UT), A cognitive approach to modeling bad news conversations
2015-13 Giuseppe Procaccianti (VU), Energy-Efficient Software
2015-12 Julie M. Birkholz (UVA), Multi-Organizational Social Network Dynamics: The Effect of Context on Social Collaboration Networks
2015-11 Yongming Luo (TUD), Designing algorithms for big graph datasets: A study of computing bisimulation and joined query processing
2015-10 Henry Herrmans (OUN), Open/C design of an integrated system to support lifelong learning
2015-09 Rodyana Klaassen (UT), HCI Perspectives on Behavior Change Support Systems
2015-08 Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions
2015-07 Maria-Kathelike Poets (UVA), Time-Aware Online Reputation Analysis
2015-06 Faheem Mehnari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes
2015-05 Christoph Büsch (U), Cryptographically Ensured Search, Pattern Holding
2015-04 Howard Spectorstra (OUN), Collaborations in Open Learning Environments
2015-03 Twan van Laarhoven (RUN), Machine learning for network data
2015-02 Faiza Bukhsh (UVA), Smart m duly: Innovative Compliance Checking in Custom Controls
2015-01 Niels Netten (UVA), Machine Learning for Relevance of Information in Crisis Response
2014-34 Shahgond Liang (UVA), Fusion and Diversity in Information Retrieval
2014-33 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
2014-32 Birgit Schnitz (OU), Mobile Games for Learning: A Pattern-Based Approach
2014-31 Paulien Meesters (UvT), Intelligent Blows: Intelligence-postured politiziering in gebiedshonden endothelium
2014-30 Kevin Vlaanderen (U), Supporting Process Improvement using Method Incrementations
2014-29 Carsten Erichoff (CWI/TUD), Contextual Multimedia Relevance Models
2014-28 Frederik Hogenboom (EUR), Automated Detection of Financial Events in News Test
2014-26 Jasmijn Marie (VU), Web Communities, Immigration and Social Capital
2014-25 Danny Plass-Oude Bos (UU), Making brain-computer interface systems better: improving usability through post-processing
2014-24 Moral Dadure (UT), Experts and Machines United Against Cyberbullying
2014-23 Yossi Ilia (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
2014-21 Christina Manteli (VU), The Effect of Generosity in Global Software Development: Analysing Transactive Memory Systems
2014-20 Tesfa Tegegne Asfaw (RUN), Service Discovery in eHealth
2014-19 Jeroen van Oijen (UVA), On Entity Resolution in Probabilistic Databases
2014-18 Leo Moors (U), Agent Technology in Agile Multithread Manufacturing and Product Support
2014-17 Peter de Kock Benamou (CWI), Anticipating Criminal Behaviour
2014-16 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
2014-15 Christiaan Mihalowicz (VU), Decentralized k-Clique Matching
2000-06 Rogier van Eijk (UU), Programming Languages for Agent Communication
2000-05 Ruud van der Pol (UM), Knowledge-based Query Formulation in Information Retrieval.
2000-04 Geert de Haan (VU), ETAG, A Formal Model of Competence Knowledge for User Interface Design
2000-03 Carolien M.T. Metselaar (UVA), Sociaal-organisatorische gevolgen van kennis Technologie; een prosesbenadering en actorperspectief.
2000-02 Koen Holtman (TUE), Prototyping of CMS Storage Management
2000-01 Frank Niessink (VU), Perspectives on Improving Software Maintenance
1999-08 Jacques H.J. Lenting (UM), Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.
1999-07 David Spelt (UT), Verification support for object database design
1999-06 Niek J.E. Wijngaards (VU), Re-design of compositional systems
1999-05 Aldo de Moor (KUB), Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems
1999-04 Jacques Penders (UM), The practical Art of Moving Physical Objects
1999-03 Don Beal (UM), The Nature of Minimax Search
1999-02 Rob Potharst (EUR), Classification using decision trees and neural nets
1999-01 Mark Slord (VU), Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products
1998-05 E.W.Oskamp (RUL), Computerondersteuning bij Straftoemeting
1998-04 Dennis Breuker (UM), Memory versus Search in Games
1998-03 Ans Steuten (TUD), A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective
1998-02 Floris Wiesman (UM), Information Retrieval by Graphically Browsing Meta-Information
1998-01 Johan van den Akker (CWI), DEGAS - An Active, Temporal Database of Autonomous Objects