MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs

Jifan Yu[1][*], **Gan Luo**[1][*], **Tong Xiao**[1], **Qingyang Zhong**[1], **Yuquan Wang**[1], **Wenzheng Feng**[1],
Junyi Luo[1], **Chenyu Wang**[1], **Lei Hou**[1,2,3], **Juanzi Li**[1,2,3][†], **Zhiyuan Liu**[1,2,3], **Jie Tang**[1,2,3]

[1]Dept. of Computer SCi.& Tech., Tsinghua University, China 100084
[2]KIRC, Institute for Artificial Intelligence, Tsinghua University, China 100084
[3]Beijing National Research Center for Information Science and Technology, China 100084
`{yujf18,luog18,zhongqy16,wangyuqu19,fwz17}@mails.tsinghua.edu.cn`
`{houlei,lijuanzi,liuzy,jietang}@tsinghua.edu.cn`

Abstract

The prosperity of Massive Open Online Courses (MOOCs) provides fodder for many NLP and AI research for education applications, e.g., course concept extraction, prerequisite relation discovery, etc. However, the publicly available datasets of MOOC are limited in size with few types of data, which hinders advanced models and novel attempts in related topics. Therefore, we present MOOCCube, a large-scale data repository of over 700 MOOC courses, 100k concepts, 8 million student behaviors with an external resource. Moreover, we conduct a prerequisite discovery task as an example application to show the potential of MOOCCube in facilitating relevant research. The data repository is now available at `http://moocdata.cn/data/MOOCCube`.

1 Introduction

Massive open online courses (MOOCs) boom swiftly in recent years and have provided convenient education for over 100 million users worldwide (Shah, 2019). As a multi-media, large-scale online interactive system, MOOC is an excellent platform for advanced application research (Volery and Lord, 2000). Since MOOC is committed to helping students learn implicit knowledge concepts from diverse courses, many efforts from NLP and AI raise topics to build novel applications for assistance. From extracting course concepts and their prerequisite relations (Pan et al., 2017b; Roy et al., 2019; Li et al., 2019) to analyzing student behaviors (Zhang et al., 2019; Feng et al., 2019), MOOC-related topics, tasks, and methods snowball in recent years.

Despite the plentiful research interests, the resource from real MOOCs is still impoverished.

Most of the publicly available datasets are designed for a specific task or method, e.g., Zhang et al.(2019) build a MOOC enrollment dataset for course recommendation and (Yu et al., 2019) is only for course concept expansion, which merely contains a subset of MOOC elements. Consequently, they are not feasible enough to support ideas that demand more types of information. Moreover, these datasets only contain a small size of specific entities or relation instances, e.g., prerequisite relation of TutorialBank (Fabbri et al., 2018) only has 794 cases, making it insufficient for advanced models (such as graph neural networks).

Therefore, we present MOOCCube, a data repository that integrates courses, concepts, student behaviors, relationships, and external resources. Compared with existing education-related datasets, MOOCCube maintains the following advantages:

• **Large-scale**: MOOCCube contains over 700 MOOC courses, 38k videos, 200k students, and 100k concepts with 300k relation instances, which provide sufficient resources for models that require large-scale data.

• **High-coverage**: Obtained from real MOOC websites and external resources, the courses, concepts, and student behaviors in MOOCCube have profuse attributes and relationships, offering comprehensive information for various related tasks.

As shown in Figure 1, a data cell of MOOC-Cube is in terms of concepts, courses, and students, which represents a learning fact, i.e., a student s learns concept k in course c. Through different queries, MOOCCube can provide various combinations of these data cells to support existing research. In this paper, we first introduce the data collection process and then give an insight into the characteristics of MOOCCube by analyzing its statistics in different aspects. We also conduct a typical NLP application task on MOOCCube and discuss the future directions on the usage of our datasets.
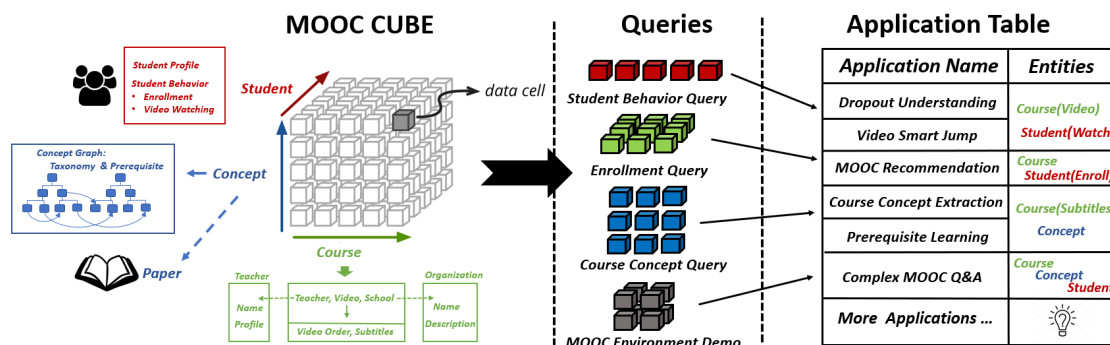
Figure 1: The framework of MOOCCube.

Our contribution is in two folds: a) an investigation of NLP and AI application research in online education, especially in MOOCs; b) a large-scale data repository of MOOCs, which organizes data in three dimensions: student behaviors, courses, and knowledge concepts.

2 Dataset Collection

2.1 An Overview of MOOCCube

Figure 1 gives an overview of MOOCCube, which models various facts of MOOCs in three main dimensions: *courses*, *concepts* and *students*. Due to the rich relationships among these entities, we organize the data into a form of a knowledge base for convenient storage and query. Through specific queries, MOOCCube can support diverse related applications, e.g., we can build a dataset for *dropout prediction* tasks by collecting a student's all behaviors in a certain course, and build a *concept extraction* dataset with all concepts in all courses. In subsequent sections, we introduce how to obtain and process the abundant data from Xue-tangX[1], one of the largest MOOC website in China, while considering the issue of privacy protection.

2.2 Course Extraction

Courses are the foundation of MOOCs and consist of a series of pre-recorded *videos*. Regarding each course as an entity, we extract the synopsis, video list, *teacher*, and the *organization*, offering this course as its attributes. As shown in Figure 1, We obtain each video's subtitle and save the order of videos for further knowledge discovery in MOOCs. Notably, we also record the description of the teacher and the organization from Wikidata[2] as an external resource.

2.3 Concept and Concept Graph

Course concepts refer to the knowledge concepts taught in the course videos. For each video, we extract 10 most representative course concepts from subtitles (Pan et al., 2017b). We also record the concept description from Wikidata and search top 10 related papers for each concept via AMiner[3] (Tang et al., 2008) as external resource. Moreover, as many NLP types of research are interested in discovering semantic relationships among concepts, we further build a novel concept taxonomy with prerequisite chains as a *concept graph* (Gordon et al., 2016).

Concept Taxonomy. A solid concept taxonomy is favorable for further research in course content (Gordon et al., 2017). However, existing taxonomies like ConceptNet (Liu and Singh, 2004) or Wiki Taxonomy (Ponzetto and Strube, 2007) cannot be directly applied to course concepts because course concepts are mostly academic terms and the non-academic categories greatly interfere with the quality of taxonomy. Thus, we select a cross-lingual term taxonomy from CNCTST[4] as a basis and lead manual annotation to build a serviceable course concept taxonomy for MOOCCube.

Prerequisite Chain. Prerequisite relation is defined as: If concept A can help understanding concept B, then there is a prerequisite relation from A to B (Gordon et al., 2016). Prerequisite relation has received much attention in recent years (Pan et al., 2017a; Fabbri et al., 2018; Li et al., 2019) and has a direct help for teaching applications. To build prerequisite chains, we first reduce the amount of candidate concept pairs by utilizing taxonomy information (Liang et al., 2015) and video dependency (Roy et al., 2019), and then lead

manual annotation. The annotation results are then employed to train different models to build a much larger distant supervised prerequisite dataset.

2.4 Student Behavior

Student behavior data not only supports relevant research (such as course recommendation (Zhang et al., 2019), video navigation (Zhang et al., 2017), dropout prediction (Feng et al., 2019)), but also indicates the relationships between courses and concepts (Liang et al., 2015). To meet different needs, we preserve the *enrollment records* and *video watch logs* of over 190,000 users from 2017 to 2019. Note that *video watch logs* record student behavior in detail, e.g., click a certain sentence, jump back to a video point, etc. Considering the data quality and privacy, we first remove the users with less than two video watching records and then anonymize the user names into UserIDs. We further shuffled these IDs and relinked them to the "most popular names"[5].

2.5 Data Processing and Annotation

We lead data processing and annotations, including 1) process the extracted course videos into subtitles; 2) process the related papers into Json files; 3) the annotation of course/video dependency; 4) large-scale annotation of concept taxonomy and prerequisite relations. All the annotations are provided by students in corresponding domains with strict quality controls[6].

3 Data Analysis

In this section, we analyze various aspects of MOOCCube to provide a deeper understanding of the dataset.

Comparison with similar datasets. Table 1 shows statistics of MOOCCube and other AI-In-Education datasets, including KDDCup2015 (Predicting dropout in MOOCs) (Cup, 2015), hierarchical MOOC recommendation (HMR) (Zhang et al., 2019), prerequisite relation learning(PRL) (Pan et al., 2017a), TutorialBank (Fabbri et al., 2018) and LectureBank (Li et al., 2019). The comparison is conducted in two aspects:

• **Data Size**. MOOCCube contains the largest data size, especially the course concept graph. For example, the number of prerequisite concept pairs

exceeds the existing datasets by almost 100 times, and hereafter supports the attempts of advanced models such as neural networks on related tasks.

• **Data Dimension**. Existing datasets are clearly divided into two categories: datasets centered on user behavior, such as HMR, they only contain very little course content information; datasets centered on course content, such as LectureBank, they focus on the concepts in the education material instead. MOOCCube organically combines these types of data in the MOOC environment so that researchers can analyze specific learning behavior.

Concept Graph. Figure 2 shows the concept distribution over different categories. Overall, we divide the concepts into 24 domains. There are significantly more concepts in engineering courses than in natural sciences or social sciences, while the number of sub-fields is the opposite. Since there are more than 1,500 valid concepts in each field, the concept information in MOOCCube is abundant. Moreover, the statistic of prerequisite concept pairs in Table 1 indicates its rarity: only 6% of concept pairs maintain a solid prerequisite relation, which explains its scarcity in existing datasets.

Student Behavior. Figure 3(a) shows the course distribution of enrolled users, which substantially fits a normal distribution. Despite a few courses with rare students, 451 courses are enrolled by over 100 users. Figure 3(b) presents a user view of the data, indicating more than 70% of users possess over ten videos watching records. These statistical results give an insight into abundant interaction between MOOCCube students, courses, and videos.

4 Application

Such a wealth of data enables MOOCCube to support multiple tasks such as *course recommendation* (Zhang et al., 2019), *concept mining* (Yu et al., 2019), etc. In this section, we conduct an important and typical task, *prerequisite relation discovery* as an example application of MOOCCube by utilizing different types of data from it. As introduced in Section 2.3, prerequisite relation indicates "what should a student learn at first". Since existing efforts have attempted to discover such relationships among concepts from different types of information, we reproduce the following methods on MOOCCube and present some basic new models.

• **MOOC-LR** and **MOOC-XG** learn such relations from the course video list and the abstracts of Wikipedia (Pan et al., 2017b), we select Logic Re-

---

[5]Published by Social Security Administration, https://www.ssa.gov/

[6]Some annotation and quality control details are in Appendix.

Dataset	Course	Video	Concept	Prerequisite	Taxonomy	Student	Enrollment	Video Watching	External Resource
KDDCup2015	39	–	–	–	–	112,448	200,904	1,319,032	–
HMR	1,302	–	–	–	–	82,535	458,454	–	–
PRL	20	1,356	573	3,504	–	–	–	–	Corpus
TutorialBank	–	–	200	794	200	–	–	–	Corpus, Paper
LectureBank	60	–	208	921	1,221	–	–	–	Corpus, Paper, Blog
MOOCCube	706	38,181	106,056	17,686	3,152	199,199	682,753	4,874,298	Corpus, Paper

Course, *Video*, *Concept*, *Student* are the sum of respective entities. *Prerequisite* is the number of relation instances, *Taxonomy* is the number of finest taxonomy categories, and *Enrollment* and *Video Watching* are the records of behavior.

Table 1: Statistics of existing NLP-in-Education datasets.
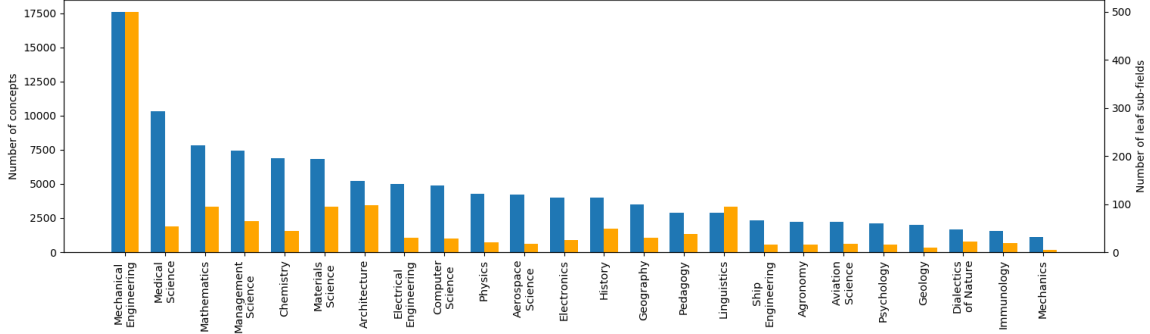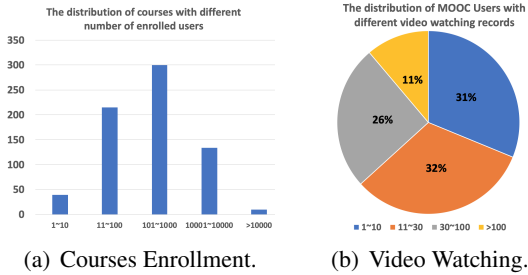

Figure 2: Concept distribution over taxonomy

(a) Courses Enrollment.   (b) Video Watching.

Figure 3: (a) shows the number of courses for different enrolled users while (b) is the user with different video watching records.

	P	R	F1-Score
MOOC-LR	0.667	0.479	0.565
MOOC-XG	0.607	0.507	0.552
PREREQ	0.606	**0.755**	**0.672**
PREREQ-S	0.651	0.730	**0.688**
PCNN	0.629	0.636	0.630
PRNN	**0.681**	0.668	0.659

Table 2: Results of prerequisite discovery.

gression and Xgboost as the classifier of the model.

• **PREREQ** employs a network to detect such relationships from course and video dependency (Roy et al., 2019). Here we present an improved version **PREREQ-S** by introducing students' video watch order to enhance the video dependency network, i.e., we sort the watched videos of each student by time and utilize these sequences for replacing the video sequences in the original paper.

• **PCNN** and **PRNN**. We present two simple DNN models, which first encode the embeddings (Cao et al., 2017) of the concept pairs and then train an MLP to classify the prerequisite ones.

Result Analysis. Overall, PREREQs perform best in F1-score, while student behavior is beneficial to

the precision of this model (PREREQ-S improves the precision to 0.651). We argue that the diverse information provided by MOOCCube helps to discover such relationships. Meanwhile, two simple DNN models perform competitive results in this task, which indicates that the existing methods are indeed limited by the amount of data (Most advanced models cannot be trained on small datasets).

5   Related Work

In this section, we introduce the research of NLP in education, especially in MOOCs, as well as several publicly available related datasets.

Existing research in MOOCs uses courses and students as the main resource, which can be divided into two categories according to the research object: one focuses on the content of the courses, such as the course concept extraction (Pan et al., 2017b), prerequisite relation discovery (Pan et al., 2017a), and course concept expansion (Yu et al.,

2019); the other focuses on the learning behavior of students, such as the prediction of dropouts (Feng et al., 2019), course recommendations (Zhang et al., 2019; Cao et al., 2019), etc. Due to the different tasks, researchers have to repeat the work to build their datasets, which arouses the original motivation of MOOCCube.

In addition, some researchers also try to obtain education information from other resources, e.g., ACL Anthology (Radev et al., 2013), TutorialBank (Fabbri et al., 2018), and LectureBank (Li et al., 2019). They collected concepts and relationships from papers and lectures and also built diverse datasets. Though they are also limited in data scale, these beneficial attempts guide the construction of MOOCCube.

6 Conclusion and Future Work

We present MOOCCube, a multi-dimensional data repository containing courses, concepts, and student activities from real MOOC websites. Obtaining large-scale data in all dimensions, MOOCCube can support new models and diverse NLP applications in MOOCs. We also conduct prerequisite relation extraction as an example application, and experimental results show the potential of such a repository. Promising future directions include: 1) utilize more types of data from MOOCCube to facilitate existing topics; 2) employ advanced models in existing tasks; 3) more innovative NLP application tasks in online education domain.

Acknowledgments

References

Yixin Cao, Lifu Huang, Heng Ji, Xu Chen, and Juanzi Li. 2017. Bridge text and knowledge by learning multi-prototype entity mention embedding. In *ACL*.

Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The world wide web conference*.

KDD Cup. 2015. Kdd cup 2015: Predicting dropouts in mooc.

Alexander Fabbri, Irene Li, Prawat Trairatvorakul, Yijiao He, Weitai Ting, Robert Tung, Caitlin Westerfield, and Dragomir Radev. 2018. Tutorialbank: A manually-collected corpus for prerequisite chains, survey extraction and resource recommendation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–620.

Wenzheng Feng, Jie Tang, and Tracy Xiao Liu. 2019. Understanding dropouts in moocs. In *Proceedings of the AAAI Conference on Artificial Intelligence, (Vol 33 No 01: AAAI-19, IAAI-19, EAAI-20)*.

Jonathan Gordon, Stephen Aguilar, Emily Sheng, and Gully Burns. 2017. Structured generation of technical reading lists. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 261–270, Copenhagen, Denmark. Association for Computational Linguistics.

Jonathan Gordon, Linhong Zhu, Aram Galstyan, Prem Natarajan, and Gully Burns. 2016. Modeling concept dependencies in a scientific corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–875, Berlin, Germany. Association for Computational Linguistics.

Irene Li, Alexander R Fabbri, Robert R Tung, and Dragomir R Radev. 2019. What should i learn first: Introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6674–6681.

Chen Liang, Zhaohui Wu, Wenyi Huang, and C Lee Giles. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674.

Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017a. Prerequisite relation learning for concepts in moocs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1447–1456.

Liangming Pan, Xiaochen Wang, Chengjiang Li, Juanzi Li, and Jie Tang. 2017b. Course concept extraction in moocs via embedding-based graph propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 875–884.

Simone Paolo Ponzetto and Michael Strube. 2007. Deriving a large scale taxonomy from wikipedia. In *AAAI*, volume 7, pages 1440–1445.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2019. Inferring concept prerequisite relations from online educational resources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9589–9594.

D Shah. 2019. Year of mooc-based degrees: A review of mooc stats and trends in 2018. class central. *Class Centrals MOOC Report*.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 990–998. ACM.

Thierry Volery and Deborah Lord. 2000. Critical success factors in online education. *International journal of educational management*, 14(5):216–223.

Jifan Yu, Chenyu Wang, Gan Luo, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2019. Course concept expansion in moocs with external knowledge and interactive game. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 4292–4302.

Han Zhang, Maosong Sun, Xiaochen Wang, Zhengyang Song, Jie Tang, and Jimeng Sun. 2017. Smart jump: Automated navigation suggestion for videos in moocs. In *Proceedings of the 26th international conference on world wide web companion*, pages 331–339. International World Wide Web Conferences Steering Committee.

Jing Zhang, Bowen Hao, Bo Chen, Cuiping Li, Hong Chen, and Jimeng Sund. 2019. Hierarchical reinforcement learning for course recommendation in moocs. In *Proceedings of the AAAI Conference on Artificial Intelligence, (Vol 33 No 01: AAAI-19, IAAI-19, EAAI-20)*.

A Data Annotation and Quality Control

As introduced in Section A, we conduct manual annotations with a quality control mechanism. Three relations need tagging: Course Dependency Chain, Concept Taxonomy, and Concept Prerequisite Chain.

• **Course Dependency Chain** is the recommended course order of learning, which is often presented by teaching assistance or mentor in school.

Many efforts for extracting prerequisite relation utilize this information (Liang et al., 2015; Roy et al., 2019). For each domain of courses, we invite three experts who have corresponding teaching experience to annotate the dependency relation among them.

• **Concept Taxonomy** annotation is in two processes: 1) For each course concept, we use a pre-trained word embedding to calculate the most likely category of it. Then three annotators in the corresponding field are asked to label whether the concept belongs to this category. 2) For the concept-category pairs that are labeled as "not belong to", we choose the brother category of the prior one as a new candidate and put the refreshed pair into the annotation pool again. Such process effectively reduces the number of invalid annotations.

• **Concept Prerequisite Chain**. To detect the prerequisite relation between concepts, we convene students in the corresponding domain as annotators. However, labeling all possible pairs is infeasible, for 100K concepts may generate over 500 billion candidate pairs. Thus we lead a distantly supervised annotation in three stages. First, we only select the concepts which occur in the same course to sample candidate concept pair. As in prior work, the annotators label if concept A is helpful to understand B. Second, we train a model as (Pan et al., 2017a) and classify other unlabeled pairs. Finally, the results with a low confidence score are labeled again to train another classifier and give all pairs a new label. This process repeats for several rounds[7], and the voting result of each pair is finally adopted. In total, 3,500 pairs are in manual labeling, and the experiments in Application use them as the test set.

Quality Control. Both of concept taxonomy and prerequisite relations are subjective (Liang et al., 2015). To prevent low-quality annotation results, we mix some golden standards (which are from existing well-organized datasets (Fabbri et al., 2018)) into the annotation pool. Once the labeling result is different from the golden standard, we lead another expert estimation to specifically confirm the truth of these conflicts and identify the annotators that can's meet the requirements.

B MOOC Q&A Dataset

Except for the data types that are introduced in the paper, we also collect and build a Q&A dataset of MOOCs, which requires an ability of language

---

[7]This process is experimentally set to 5 rounds.

Course Name	Number of concepts included
Data and Structure 1	140
Data and Structure 2	117
Network Technology and Applications	125
The Basics of Programming in C++ Language	84
The Advanced Design of C++ Program	77
Introduction to Computer Science and Python Programming	116
Operating System	143
Java Programming Design	62
Artificial Intelligence	66
Artificial Intelligence for Beginner	54
5G and Artificial Intelligence	71
Big Data and Machine Learning	236

Table 3: Course name list.

	Type	One-hop	Multu-hop	Total	All Types
MOOCQA Dataset	Query(Type A)	5,504	10,615	16,119	
	Judge(Type B)	16,324	14,301	30,625	53,311
	Count(Type C)	3,384	3,183	6,567	

Table 4: Statistics of questions.

Entity/Relationship	Table Name	Number of Rows
Entity	concept	700
	course	12
	paper	5,927
	school	208
	teacher	1,733
	user	4,723
	video	1,242
Relationship	concept_field	44
	concept_paper	5,927
	course_concept	10,346
	course_video	1,591
	school_course	705
	school_teacher	2,130
	teacher_course	2,349
	user_course	24,933
	video_concept	4,040

Table 5: Statistics of entities and relationships in MOOC Q&A.

understanding and multi-hop reasoning , to provide a comprehensive resource for more possible applications of MOOCs.

Here are the methods we followed to collect the QA dataset. We divide the dataset into one-hop questions and multi-hop questions. An one-hop question only involves a single head entity and a single predicate in the knowledge, while a multi-hop question may contain several entities and to answer the question needs to reason over several facts in the knowledge graph.

We design 22 types of 1-hop question schema and 20 types of multi-hop question schema based on the meaningful real queries we collected from MOOC platform. Each schema is paraphrased into 4 different templates and questions are generated by random sampling from the text template pool. Triples related with twelve typical courses are used in case that the model wont run out of memory. The twelve courses are listed as Table 3.

The twelve courses are all from the computer science field. They cover different levels of courses in computer science and the internal prerequisite-successive relationships between the twelve courses typically represent the real relations between courses in MOOC platform. The model trained on our dataset is expected to provide MOOC users with information and further related knowledge they need. The type and number of entities and relationships are shown in Table 5.

Besides, to make our dataset closer to the actual scenario, three types of questions are contained in MOOCQA Dataset, which are Query, Judge and Count. When answering Query questions, model is expected to offer the correct entities in knowledge graph. As for Count questions, the count of the related entities is required. For Judge questions, the model should make a clear judgement of the

factoid description in the question.

In MOOCQA Dataset, each line is a question sample. In addtion to the question and its corresponding answer, we provide more information including entity ids, question type, etc. Question, supporting fact and answer are separated by "\t". If the answer consist of several entities, they will be separated by '|'.