MOOC-Rec: Instructional Video Clip Recommendation for MOOC Forum Questions*

Peide Zhu
Delft University of Technology
p.zhu-1@tudelft.nl

Jie Yang
Delft University of Technology
j.yang-3@tudelft.nl

Claudia Hauff
Delft University of Technology
c.hauff@tudelft.nl

ABSTRACT

In this work, we address the information overload issue that learners in Massive Open Online Courses (MOOCs) face when attempting to close their knowledge gaps via the use of MOOC discussion forums. To this end, we investigate the recommendation of one-minute-resolution video clips given the textual similarity between the clips' transcripts and MOOC discussion forum entries. We first create a large-scale dataset from Khan Academy video transcripts and their forum discussions. We then investigate the effectiveness of applying pre-trained transformers-based neural retrieval models to rank video clips in response to a forum discussion. The retrieval models are trained with supervised learning and distant supervision to effectively leverage the unlabeled data—which accounts for more than 80% of all available data. Our experimental results demonstrate that the proposed method is effective for this task, by outperforming a standard baseline by **0.208** on the absolute change in terms of precision.

Keywords

MOOC, Discussion Forum, Video Clip Transcripts, Clip Recommendation

1. INTRODUCTION

Massive Open Online Courses (MOOCs) provide open access to world class courses for the public, which greatly improves the opportunities in online learning. The discussion forum is a major component of a MOOC as it is the primary communication tool among learners and instructors [1] to moderate the lack of physical access in MOOCs. It can help learners build a sense of belonging and learn from peers, or help instructors monitor learner affect and academic progress [2]. However, since questions targeting the same video content are scattered among discussion threads, without supporting navigation facilities, learners cannot effectively retrieve valuable discussions for a particular piece of content. In addition, learners' posts seeking help may be drowned out by the many other competing posts, making it hard for learners to get attention from instructors and peers. The unstructured, unorganized forums with a large amount of discussions (that can lead to information overload [19]) are hindering instructors and learners to benefit from them, decrease community interaction, reduce responsiveness in forums and in the end lead to low MOOC retention rates [20, 13].

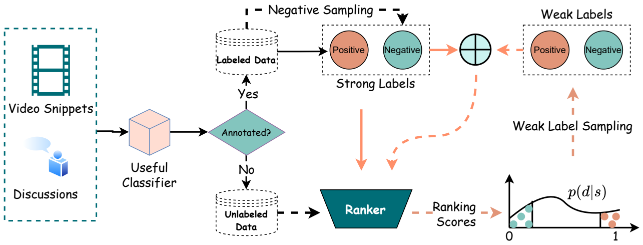

Figure 1: Overview of MOOC-Rec.

Existing works directed at addressing the information overload issue in MOOC forums have proposed more effective navigation tools to identify instructional video contents and make recommendations of a ranked list of video clips. For example, [2] classify posts that need help and employ bag-of-words based retrieval techniques to map those posts to minute-resolution course video clips. The clip recommendation algorithm is evaluated on posts from one course. [17] built a recommender system to generate a ranked list of video clips giving a student's question with a deep neural network; they evaluate the system with 50 questions. Despite these attempts, we argue that prior works on video clip recommendation suffer from a lack of training data, and as a consequence report evaluations only on small-scale data. It remains a challenge to develop and evaluate a system that can scale to thousands of MOOCs, across different domains.

In our work, we first address the lack of training data issue by creating `MOOC-CLIP`, a novel large-scale dataset from Khan Academy [1], that includes video transcripts and forum posts (both questions and answers) using raw data available from `LearningQ` [3], an open source tool and dataset for educational question generation. Second, we propose `MOOC-Rec`,

*The first author has been supported by the China Scholarships Council (CSC). The last author has been supported by NWO project SearchX (639.022.722) and NWO Aspasia (015.013.027).

P. Zhu, C. Hauff, and J. Yang. MOOC-Rec: Instructional video clip recommendation for MOOC forum questions. In A. Mitrovic and N. Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 705–709, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.6853055

[1] `https://www.khanacademy.org/`

a dense retrieval based instructional video clip recommendation system for MOOC forum questions. For each content-related thread, MOOC-Rec recommends a ranked list of video clips that are likely relevant and helpful for answering the question. Although dense retrievers have been applied in various retrieval tasks such as DPR [6] and ColBERT [7], it is unknown whether they are an effective approach for MOOC video clip recommendation. Lastly we point out that only 11.57% of all discussions in our dataset are labeled with a target video clip, which poses challenges for training MOOC-Rec with limited labeled data and abundant unlabeled resources.

We here first investigate the effectiveness of MOOC-Rec and then we address the scarcity of labeled data by using distant supervision and in-batch negatives to train the ranker. The comprehensive experiments on our large-scale dataset which consists of about 274K discussions show that our systems significantly improve the clip recommendation performance by outperforming a standard baseline by 0.208 in terms of precision.

2. THE MOOC-CLIP DATASET

To address the lack of research data, we create a large-scale dataset using raw data crawled with LearningQ[2] from Khan Academy, a MOOC platform which allows learners to ask and answer questions about the learning materials during learning. We keep video transcripts, forum questions and answers of MOOCs which have both transcripts and discussions available.

Learners use discussion forums in different ways. Besides asking questions related to the course materials, they may also discuss irrelevant topics [14] for the purposes of socializing, spamming, or expressing their appreciation for the course materials. Some questions posted by learners also suffer from a lack of proper context, or are too generic. Therefore, it is necessary to remove these relatively—for our purposes—low-quality questions. In line with LearningQ, we consider a user-generated question to be 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; (ii) the question is context-complete, containing sufficient context information to enable other learners to answer the question; and (iii) the question is not generic. Besides labeled questions in LearningQ, we manually labeled 2K questions among other topics. We also labeled 5K questions based on their lexical relevance to video transcripts (2.5K with highest BM25 scores as useful, 2.5K with lowest BM25 scores as negative) in order to exclude non-relevant questions. In total, there are 13,290 labeled questions over 8 topics. We found 60.9% of them to be *useful* and 39.1% of them to *not be useful.* We keep all items belonging to 3 topics (2,344 in total) as unknown set for our cross-topics evaluation, 8,766 questions on the remaining 5 topics for training, and 2,186 questions as known topic test set. We train a BERT-based text sequence classifier for useful question classification. Table 1 summarizes its performance.

	Same Topic			Cross-Topics		
Method	Acc	Rc	F1	Acc	Rc	F1
Q	89.40	96.68	92.90	77.20	74.49	75.82
Q+C	89.75	96.54	93.02	73.30	82.68	77.71

Table 1: Useful question classifier results.

During preprocessing, we first remove noisy discussions which contain only meaningless tokens, as well as videos which have no discussions. Then we apply the useful question classifier on all items(522K) and retrain only items are classified as useful. In the end, we retain 273,887 discussions from 7,349 videos of 6 topics.We use regular expressions to retrieve discussions where learners label posts with exact timestamps in questions or answers. We split the video transcripts into snippets with a one minute length. The discussions and the snippets which cover the timestamp are labeled as *positive* items. The other discussions are treated as unlabeled. Table 2 and Figure 2 summarize the data statistics. In summary, there are 31,680 positive labeled items and 240,551 unlabeled items, i.e. 11.57% of all discussions are labeled.

Split	**#V**	**#S/V**	**#W/S**	**#W/Q**	**#W/A**
Train	4590	7.91	198.51	39.96	80.89
Dev	895	8.37	199.04	40.02	79.26
Test	1126	8.14	198.64	39.67	81.92
Unlabeled	7283	7.70	197.96	38.46	78.58

Table 2: Dataset overview, in terms of videos (#V), snippets (#S) per video, discussions (#D) per video, clip (#W), the number of words per question (Q) and the number of words per answer (A)

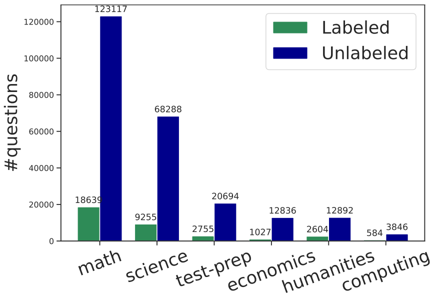

Figure 2: Dataset overview regarding the number of labeled and unlabeled questions in each topic. We can see the unbalanced distribution questions in each topic.

This dataset also covers a series of educational topics including math, science, careers, humanties, etc. We conduct an exploratory analysis along each topic dimension which is shown in Fig 2. We observe a topic imbalance, e.g. discussions under math and science topics account for 78.88% of labeled items and 76.82% of all items. The labeled data is then split into 80% and 20% for training and test sets respectively based on the number of discussions in each set.

3. METHODOLOGY

The problem of MOOC video clip recommendation studied in this paper can be described as follows. Given a forum

[2] https://github.com/AngusGLChen/LearningQ

discussion question, the system retrieves a ranked list of the most relevant video clips as represented by their transcripts. We assume the questions filtered by the useful question classifier are relevant to the course materials, and the most relevant video clips should be instructional for learners. Assume a MOOC video $\mathcal{V}$ lasts for T seconds, then we split it into s $t = 60$ seconds clips, where $s = \lceil \frac{T}{t} \rceil$. Then the video $\mathcal{C}$ contains clips $c_1, c_2, \cdots, c_s$. Each clip c_i is represented with its transcripts, which can be viewed as a sequence of tokens $w_1^i, w_2^i, \cdots, w_{|c_i|}^i$. We also formally define a discussion as $d_i = [q_i, \{a_i\}]$, where $\{a_i\}$ are the answers to the question q_i. Note that in some cases the question has not been answered yet, which is common in MOOC forums. The task is retrieve a ranked list of clips $c_{i,1}, c_{i,2}, \cdots, c_{i,s}$ given each discussion d_i. Notice that the video clip recommender needs to work effectively for MOOCs in different domains that the corpus covers. Formally speaking, the recommender $\mathcal{R} : (d, \mathbf{C}) \rightarrow \mathbf{C}_{\mathcal{R}}$ is a function that takes a discussion d and video clip list $\mathbf{C}$ as the input and returns a ranked list of clips $\mathbf{C}_{\mathcal{R}}$. We can also choose to only return the top-K most relevant clips.

3.1 Dual-Encoder

We employ a standard neural IR architecture [6] for the ranker. It uses a dense encoder $E_C(\cdot)$ which encodes the video clip transcripts into m-dimensional real-valued vectors. At run-time, `MOOC-Rec` maps the input discussion $d = [q, a]$ to another m-dimensional vector using the query encoder $E_Q(\cdot)$, and retrieves the top-k most closest video clip vectors from the same video. We use cosine similarity to model the similarity between the discussion and the clip vectors by the following function:

$$sim(d, c) = \cos(E_Q(d), E_C(c)). \tag{1}$$

The goal of training is to learn a better embedding function for both the clips and discussions which can map relevant pairs of discussions and clips to vectors with smaller distance, i.e. higher similarity, so that the similarity function $sim(d, c)$ becomes a good ranking function for the task of MOOC video clip recommendation. This is essentially a *metric learning* problem [9, 11, 6].

Let $\mathcal{M} = \{\langle d_i, c_i^+, c_{i,1}^-, \cdots, c_{i,n}^- \rangle\}_{i=1}^m$ be the training MOOC discussion corpus that contains m instances. Each example has one discussion $d_i = [q_i, a_i]$, one relevant (positive) video clip transcript c_i^+, and n irrelevant (negative) clips $c_{i,j}^-$. We train the retrieval model by optimizing the negative log likelihood of the positive clip:

$$L(d_i, c_i^+, c_{i,1}^-, \cdots, c_{i,n}^-) = -\log \frac{e^{sim(d_i, c_i^+)}}{e^{sim(d_i, c_i^+)} + \sum_{j=1}^n e^{sim(d_i, c_{i,j}^-)}}$$

Positive and Negative Video Clips. For labeled discussions, positive and negative video examples are explicit. We use the video clip whose time duration contains the timestamp of the discussion as the positive example. All other video clips *from the same video* can be treated as negatives. As MOOC videos vary in the number of clips and to boost the model training and balance the number of positive and negative examples, we selected n of them as the training negative examples. We apply in-batch negatives [5, 6] for training. In this case, the positive clips for other questions are also treated as the negatives for the current question.

Distant Supervision with Unlabeled Data. As we show in Table 2, over 80% of all discussions are unlabeled (i.e. there is no video timestamp available). It would be labor-intensive and expensive to create human annotations. Thus, we adopt *distant supervision* [10] to effectively utilize the rich unlabeled data and train a better model with them. This process involves training the model with noisy weakly labeled data. `MOOC-Rec` is able to achieve over 50% precision in top-1 prediction and over 70% in top-3 with a Recall@3 of over 80%. Therefore, we use the ranker trained on the labeled training set as the scorer and clips with the highest $sim(d, c)$ are selected as positives while the clips with the lowest $sim(d, c)$ (besides top-3) as negatives. The weakly labeled data are then used to train the ranker.

Inference. During inference time, we pre-compute all clip embedding v_c by applying the clip encoder E_C to all MOOC video clips offline. Given a discussion $d = [q, a]$ at run-time, we concatenate the question and answers if a is available and compute the discussion embedding $v_d = E_Q(d)$. The clips are then ranked by $sim(d, c)$ and the top-k are retrieved.

Although encoders can be implemented in many different ways [10], in this work, we use two independent **BERT** [4] variant models as encoders and the mean value of all token embeddings is used as the final representation. We tokenize clip transcripts and truncate the token list to maximum length of 512 (starting with `[CLS]` and ending with the `[SEP]` token). The discussion encoder works as a *query* encoder in typical neural IR systems. Instead of using separate encoders for questions and answers of the discussion, in our design both of them share the same encoder. In this way, we train a better query encoder for questions by taking advantage of important answer information.

3.2 Cross-Encoder

Both the cross-encoder and dual-encoder are two common approaches for matching sentence pairs. While the dual-encoder produces sentence embedding vectors for clips and discussions independently, the cross-encoder treats the clip recommendation for discussions as a sequence classification task and performs full self-attention over the entire sequence. We concatenate the video clip transcripts and the discussions (question and answers) with the `[SEP]` token as the input to the transformer network. The `[CLS]` token embedding is then passed to a binary classifier to predict the binary relevance between them.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Settings

Implementation. Two BERT variants: MPNet [16] (*abbrv.* `MP`, embedding size: 768) and MiniLM [18] (*abbrv.* `MP`, embedding size: 384) are used as text encoders. We implement dual-encoders using pre-trained weights provided by

Sentence-Transformers library [3] [15]. Both models are pre-trained on a large and diverse dataset of over 1 billion training query-paragraphs pairs for the semantic search task. The Adam optimizer [8] with warming-up and cosine schedule is used for training; we set the maximum learning rate to $lr = 2e^{-5}$, $\epsilon = 1e^{-8}$ and the warmup steps to 1000. For the **cross-encoder** baseline, we follow previous research [10, 12]. The **BM25** baseline is based on the Okapi BM25 implementation of the `rank_bm25` library [4]. We train our models using 8 GTX-1080 GPUs for 10 iterations with a batch size of 32. As Figure 3 shows, after one iteration, both clip recommendation systems outperform the BM25 baseline.

Table 3: Performance of the proposed `MOOC-Rec` ranker and baselines on the test set in terms of rank-aware metrics. MLM/MP$_{\text{dual}}$ represents the `MiniLM` or `MPNet` based dual-encoder and MLM/MP$_{\text{cross}}$ represents the `MiniLM` or `MPNet` based cross-encoder. "PT" represents ranker performance using pre-trained encoders without fine-tuning. "FT" means fine-tuned model performance. "WL" means the model performance after training with weakly labeled data.

Method		P@1	MRR	MRR@3	nDCG	nDCG@3
	BM25	0.417	0.600	0.550	0.696	0.593
PT	MLM$_{\text{cross}}$	0.132	0.346	0.254	0.497	0.297
	MLM$_{\text{dual}}$	0.422	0.614	0.568	0.707	0.617
	MP$_{\text{cross}}$	0.135	0.344	0.248	0.495	0.288
	MP$_{\text{dual}}$	0.386	0.583	0.529	0.683	0.576
FT	MLM$_{\text{cross}}$	0.511	0.677	0.641	0.755	0.683
	MLM$_{\text{dual}}$	0.529	0.692	0.658	0.767	0.700
	MP$_{\text{cross}}$	0.613	0.745	0.716	0.807	0.750
	MP$_{\text{dual}}$	0.570	0.720	0.690	0.788	0.730
WL	MLM$_{\text{cross}}$	0.540	0.696	0.661	0.770	0.700
	MLM$_{\text{dual}}$	0.520	0.683	0.646	0.760	0.687
	MP$_{\text{cross}}$	**0.625**	**0.751**	**0.722**	**0.812**	**0.754**
	MP$_{\text{dual}}$	0.557	0.711	0.680	0.782	0.720

4.2 Effectiveness of Dense Retrieval

Performance Comparison with Baseline. After several iterations, the models' performance first improves gradually and then becomes steady as illustrated in Figure 3, which shows the effectiveness of the training system and the effectiveness of the proposed models. Table 3 summarizes the models' effectiveness on the test set. We use BM25 as our baseline. Sparse vector-space models and the probabilistic BM25 model have been widely used in instructional clip recommendation systems. BM25's effectiveness in terms of Precision@1 (P@1) and MRR is 0.417 and 0.60 respectively, which shows queries possess more lexical similarity to related MOOC clips than other clips in the course video and BM25 is an effective and strong baseline for this task. First, we find that without fine-tuning, the pre-trained dual-encoder can achieve similar (MPNet), or even better (MiniLM-L6) performance than the BM25 baseline, while the cross-encoders cannot make clip recommendation for discussions without training. Second, we observe significant gains ($p = 1.95e^{-7}$) when using the `MOOC-Rec` neural ranker after it has been trained on the data, with gains of over 0.15 in P@1 and over 0.19 in nDCG scores compared to the BM25 baseline. Thus, dense retrieval is an effective instructional MOOC clip recommendation approach for forum discussions which can model the relevance between discussions and clip transcripts.

Impact of Model Size. To compare the impacts of model size, we use one distilled transformer model `MiniLM` which contains **22M** parameters and one BERT size model `MPNet` which contains **109M** parameters. As Table 3 shows, in both cross-encoder and dual-encoder settings, the larger model (i.e. `MPNet`) achieves better effectiveness after training, which shows that the transformer model with more parameters may have a better potential to model the relevance between clips and discussions.

Comparison of Cross-Encoder and Dual-Encoder. Both cross-encoder and dual-encoder are commonly used for sentence pair matching problems. In Table 3, we observe that with the distilled transformer model the dual-encoder outperforms the cross-encoder by 0.018 in terms P@1. However, with large model, the cross-encoder outperforms dual-encoder by 0.043 on P@1, and around 0.02 on other metrics. Despite the performance advantage of the cross-encoder with a large model, as outlined in Section 3.2, we observe a massive computational overhead with the cross-encoder as illustrated in Figure 5.

Effect of Distant Supervision. In the weakly-labeled data (WL) section of Table 3, we summarize the different models' performance after distant training with weakly labeled data. Compared with model trained with labeled data only, cross-encoders benefit from WL (+0.029 for `MiniLM` and +0.012 for `MPNet` in terms P@1), while dual-encoders perform gets worse (-0.009 for `MiniLM` and -0.013 for `MPNet` in terms P@1). Our hypothesis is that although `MOOC-Rec` achieves a good effectiveness after the initial training, the weakly labeled data created with it still contains considerable noisy content.

5. CONCLUSIONS

We studied the task of video clip recommendation in the context of MOOC forums which has the eventual goal to reduce learners' information overload. We created a novel dataset `MOOC-Clip` which includes video transcripts and discussions. We systematically investigated how well the state-of-art pre-trained neural IR models work for the task of MOOC clip recommendation, and proposed a framework including data preparation, useful question classification, clip ranker and weak supervision training for this task. We conducted the experiments with both cross-encoders and dual-encoders. The results on our dataset show that neural IR approaches are indeed effective—at the same time, a P@1 value of less than 0.63 (at best) shows that we are still far away from solving this task. In future work, we plan to further investigate the factors that affect `MOOC-Rec`'s effectiveness such as the clip duration and methods of creating weak labels.

[3] `https://github.com/UKPLab/sentence-transformers`

[4] `https://github.com/dorianbrown/rank_bm25`

References

[1] P. Adamopoulos. What makes a great mooc? an interdisciplinary analysis of student retention in online courses. 2013.

[2] A. Agrawal, J. Venkatraman, S. Leonard, and A. Paepcke. Youedu: addressing confusion in mooc discussion forums by recommending instructional video clips. 2015.

[3] G. Chen, J. Yang, C. Hauff, and G.-J. Houben. Learningq: a large-scale dataset for educational question generation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] M. Henderson, R. Al-Rfou, B. Strope, Y.-H. Sung, L. Lukács, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017.

[6] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

[7] O. Khattab and M. Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[9] B. Kulis et al. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013.

[10] J. Lin, R. Nogueira, and A. Yates. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*, 2020.

[11] B. Mitra, N. Craswell, et al. *An introduction to neural information retrieval*. Now Foundations and Trends, 2018.

[12] R. Nogueira and K. Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.

[13] A. Ntourmas, N. Avouris, S. Daskalaki, and Y. Dimitriadis. Evaluation of a massive online course forum: design issues and their impact on learners' support. In *IFIP conference on human-computer interaction*, pages 197–206. Springer, 2019.

[14] A. Ntourmas, S. Daskalaki, Y. Dimitriadis, and N. Avouris. Classifying mooc forum posts using corpora semantic similarities: a study on transferability across different courses. *Neural Computing and Applications*, pages 1–15, 2021.

[15] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[16] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

[17] P. Trirat, S. Noree, and M. Y. Yi. Intellimooc: Intelligent online learning framework for mooc platforms. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pages 682–685, 2020.

[18] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[19] D. A. Wiley and E. K. Edwards. Online self-organizing social systems: The decentralized future of online learning. *Quarterly review of distance education*, 3(1):33–46, 2002.

[20] D. Yang, M. Wen, I. Howley, R. Kraut, and C. Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 121–130, 2015.

APPENDIX

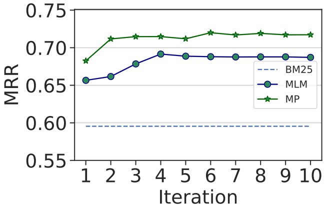

Figure 3: System performance along each training iteration.

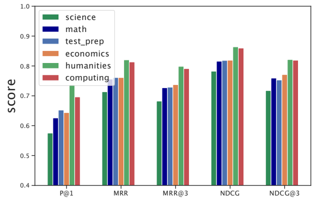

Figure 4: Performance along different topics.

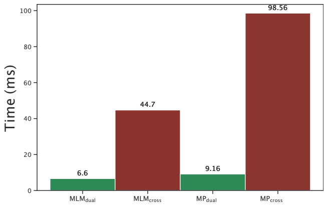

Figure 5: Average Processing Time.