A Novel Cascade Model for Learning Latent Similarity from Heterogeneous Sequential Data of MOOC

Zhuoxuan Jiang[1] and **Shanshan Feng**[2] and **Gao Cong**[2] and **Chunyan Miao**[3] and **Xiaoming Li**[1]
[1]School of Electronics Engineering and Computer Science, Peking University, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore
{jzhx,lxm}@pku.edu.cn, {sfeng003@e.,gaocong@,ascymiao@}ntu.edu.sg

Abstract

Recent years have witnessed the proliferation of Massive Open Online Courses (MOOCs). With massive learners being offered MOOCs, there is a demand that the forum contents within MOOCs need to be classified in order to facilitate both learners and instructors. Therefore we investigate a significant application, which is to associate forum threads to subtitles of video clips. This task can be regarded as a document ranking problem, and the key is how to learn a distinguishable text representation from word sequences and learners' behavior sequences. In this paper, we propose a novel cascade model, which can capture both the latent semantics and latent similarity by modeling MOOC data. Experimental results on two real-world datasets demonstrate that our textual representation outperforms state-of-the-art unsupervised counterparts for the application.

1 Introduction

With the rapid development of Massive Open Online Courses (MOOCs), more and more learners participate in MOOCs (Anderson et al., 2014). Due to the lack of effective management, most of the discussion forums within MOOCs are overloaded and in chaos (Huang et al., 2014). Therefore, a key problem is how to manage the forum contents.

To manage the forum contents, threads of forums can be regarded as documents and be classified to groups. There are several straightforward methods, such as defining sub-forums according to weeks and asking learners to tag threads. However their effectiveness is limited (Rossi and Gnawali, 2014), because learners have few incentives to tag threads. Recently, machine learning solutions have been proposed, e.g., content-related thread identification (Wise et al., 2016), confusion classification (Agrawal et al., 2015) and sentiment classification (Ramesh et al., 2015). However they are developed for specific research problems and cannot be applied to our problem. Moreover, they require labeling data which needs domain experts to label data for different courses.

We observe that the video clips of a MOOC would have many well-formed subtitles composed by instructors. Moreover, within MOOC settings, the course contents can be broken down to knowledge points, and each video clip just corresponds to a knowledge point. Consequently, we propose to fulfill the application, which is to associate threads to subtitles of video clips, i.e., thread-subtitle matching. By this way, the relevant videos to the threads can be recommended to learners, and the chaotic threads in discussion forums can also be well grouped.

However, it is challenging to identify the relevant video clips for threads without labeling data. To address this issue, we regard it as a document ranking problem based on the calculation of similarity between documents. The key problem of this task is to learn a textual representation, which can cluster similar documents and meanwhile distinguish irrelevant ones.

Intuitively, Bag-of-words model (BOW) can be utilized to calculate the similarity between threads and subtitles (Salton and Buckley, 1988). However, BOW cannot effectively capture semantics of words and documents. In addition, recently-studied semantic word embeddings, e.g., Word2Vec (Mikolov et al., 2013), can capture the semantics. Para2Vec (Le and Mikolov, 2014) can capture the similarity to some degree, but not explicitly model the latent similarity of documents.

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2768–2773
Copenhagen, Denmark, September 7–11, 2017.

Since the latent similarity is crucial to determine whether a document can be associated to the right target, in our task, the document representation is expected to preserve both the latent semantics and similarity.

In this paper, we leverage two kinds of sequential information: 1) word sequence of subtitles and forum contents, and 2) clickstream log of learning behaviors. Specifically, different from conventional representation learning tasks, e.g. Word2Vec and Para2Vec, we consider the clickstream data, which reflects the relationship between thread and video's subtitle. For instance, if a user watches a video and then clicks a thread in forums, the video would be relevant to the thread. In order to learn representations from the two types of data, we propose a novel cascade model.

Our basic idea is to jointly model three components: 1) word-word coherence, 2) document-document coherence, and 3) word-document coherence. The three components are cascaded for learning the low-dimensional word embeddings. Then the learned embeddings are used to calculate similarities between threads and subtitles.

To summarize, our contributions include:

- We study an application-oriented research problem, which is how to capture the latent similarity when learning text representation.
- We propose a novel cascade model to learn the document representation from heterogeneous sequential data: 1) word sequence and 2) learners' clickstream.
- We collect two real-world MOOC datasets and conduct thorough experiments. The results demonstrate that our proposed model outperforms the state-of-the-art unsupervised counterparts on the application.

2 Related Work

MOOC data has attracted extensive research attention and many interesting research problems have been studied. For example, dropout predicting (Qiu et al., 2016), sentiment analysis of learning gains (Ramesh et al., 2015), instructor intervention (Chaturvedi et al., 2014) and answer recommendation (Jenders et al., 2016), etc. Particularly, (Agrawal et al., 2015) considers a similar task as ours, which is to recommend video clips to threads. But its solution is designed for the specific task and needs labeling data. Our solution is an unsupervised learning method and the learned embeddings have other applications, e.g. thread retrieval.

How to represent text is a fundamental research problem in the field of information retrieval. Existing approaches can be generally classified into unsupervised methods and supervised methods (Tang et al., 2015). Although supervised embeddings can obtain good performance in specific tasks, such as using deep neural network (Mikolov et al., 2010; Kim, 2014), they need human efforts to get labels. Unsupervised word embeddings usually leverage various levels of textual information. For example, Word2Vec learns word embeddings based on word coherence. Para2Vec utilizes word and document coherence to learn their embeddings. Particularly, Hierarchical Document Vector (HDV) (Djuric et al., 2015) leverages both streaming documents and their contents to achieve better representation, which is similar to our proposed model. However, HDV regards the documents as the context of words, which cannot learn the latent similarity, since it fails to explicitly reflect the relationship between document and word. In order to model the heterogeneous MOOC data, we develop a cascade representation model. To our knowledge, (Jiang et al., 2017) also proposes an unsupervised learning model (called NOSE) for the task of thread-subtitle matching within MOOC settings. However, NOSE needs to build a heterogeneous textual network beforehand and may suffer from heterogeneous issue, which our model can avoid.

Recently, representation learning has been applied to many tasks, such as network embedding (Grover and Leskovec, 2016) and location embedding (Feng et al., 2017). In this paper, we focus on learning representation of words and documents in MOOCs.

3 Cascade Model

Based on our observation, we utilize two kinds of sequential information: 1) word sequences of subtitles and threads, and 2) clickstream of learning behaviors. In this paper, we regard the subtitles or threads consistently as *documents*. Particularly, we discover that the log of learners' clickstream, i.e., the click records of watching videos, reading threads and posting threads in a chronological order, can reflect the document-level latent semantics. An intuitive explanation is that a learner who

jumps from videos to threads may look for further relevant information from forums when s/he is watching a video, or s/he wants to review the relevant videos when s/he reads a thread.

However, learning from the log of clickstream data merely guarantees that similar documents are close enough in the embedding space, while different documents cannot be scattered. To address this issue, we attempt to strengthen the relationship between words and their affiliated documents. Thus, words within the same documents would be gathered and otherwise scattered in the embedding space. Consequently, the latent similarity can be embodied by word embeddings.

Based on the aforementioned idea, we can model the data by three components: 1) latent semantics at word level, 2) latent similarity at document level, and 3) latent similarity between words and documents. To integrate all the three kinds of information into a uniform learning framework, we propose a novel cascade model, as shown in Fig. 1. $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$ correspond to the log-likelihood of three components respectively. Formally, we aim at minimizing the log-likelihood function:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \tag{1}$$

Note that $\mathcal{L}_3$ not only learns the latent similarity, but also builds a connection between words and documents. In this way, our learned word embeddings can be adopted to our task without learning classifiers by labeling data.

3.1 Word-level Latent Semantics

As to the part of $\mathcal{L}_1$, corresponding to the red/bottom part of Fig. 1, we leverage the Word2Vec model to learn the semantics of words. In this paper, we take the Continuous bag-of-words (CBOW) architecture. The objective function is to minimize the log-likelihood:

$$\mathcal{L}_1 = -\sum_{t=1}^{T} \log \mathbb{P}(w_t | w_{t-c_w} : w_{t+c_w}) \tag{2}$$

where c_w is the context window length used in word sequence, and $w_{t-c_w} : w_{t+c_w}$ is the subsequence $(w_{t-c_w}, \ldots, w_{t+c_w})$ excluding w_t itself. The probability $\mathbb{P}(w_t | w_{t-c_w} : w_{t+c_w})$ is defined by the softmax function $\frac{\exp(\bar{\mathbf{v}}^T \mathbf{v}_{w_t})}{\sum_{w=1}^{W} \exp(\bar{\mathbf{v}}^T \mathbf{v}_w)}$, where $\mathbf{v}_{w_t}$ is the vector representation of word w_t, and $\bar{\mathbf{v}}$ is averaged vector representation of the subsequence. Two methods can be employed to calculating $\mathcal{L}_1$: hierarchical softmax and negative sampling (Mikolov et al., 2013).

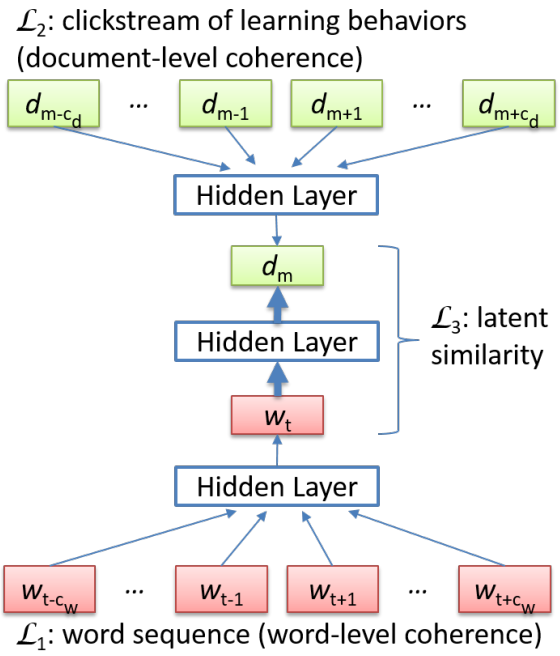


Figure 1: The architecture of proposed model which is cascaded by three parts: $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$.

3.2 Document-level Latent Similarity

Similar to $\mathcal{L}_1$, we adopt the CBOW architecture for calculating $\mathcal{L}_2$, as shown by the green/top part of Fig. 1. The objective function is to minimize the log-likelihood:

$$\mathcal{L}_2 = -\sum_{m=1}^{M} \log \mathbb{P}(d_m | d_{m-c_d} : d_{m+c_d}) \tag{3}$$

where M is the number of documents, c_d is the context window length used in clickstreams, and $d_{m-c_d} : d_{m+c_d}$ is the sub-sequence $(d_{m-c_d}, \ldots, d_{m+c_d})$ excluding d_m itself. The probability $\mathbb{P}(d_m | d_{m-c_d} : d_{m+c_d})$ is also the softmax function. Methods of hierarchical softmax and negative sampling can be employed to approximate the log-likelihood function.

3.3 Document-Word Latent Similarity

To learn the latent similarity, we make use of the relationship between words and documents, and then similar documents can be clustered, while different documents are scattered. Therefore, we propose the third component, $\mathcal{L}_3$, shown in the middle part of Fig. 1. Different from $\mathcal{L}_1$ and $\mathcal{L}_2$, we employ negative sampling of documents

to calculate its approximation, because there are numerous threads in MOOC forums. Given a pair (w_t, d_m), representing that word w_t appears in document d_m, $\mathcal{L}_3$ is denoted as:

$$\mathcal{L}_3 = \sum_{w_t} \left(-\log \sigma(\mathbf{v}_{d_m}^T \mathbf{v}_{w_t}) + \sum_{c=1}^{C} \log \sigma(\mathbf{v}_{d_c}^T \mathbf{v}_{w_t}) \right) \tag{4}$$

where $\sigma(x)$ is the sigmoid function and C is the number of sampled negative documents.

3.4 Model Training

We adopt stochastic gradient descent (SGD) to minimize $\mathcal{L}$. As to the components of $\mathcal{L}_1$ and $\mathcal{L}_2$, we exploit the training methods proposed in (Mikolov et al., 2013) to the two kinds of sequences, i.e., words and documents, respectively. For training $\mathcal{L}_3$, given the pair (w_t, d_m), we calculate the gradients:

$$\frac{\partial \mathcal{L}_3}{\partial \mathbf{v}_{d_j}} = \left(\sigma(\mathbf{v}_{d_j}^T \mathbf{v}_{w_t}) - \mathbb{1}(j = m) \right) \mathbf{v}_{w_t}, \tag{5}$$

$$\frac{\partial \mathcal{L}_3}{\partial \mathbf{v}_{w_t}} = \left(\sigma(\mathbf{v}_{d_j}^T \mathbf{v}_{w_t}) - \mathbb{1}(j = m) \right) \mathbf{v}_{d_j}, \tag{6}$$

where d_j represents both the positive and negative samples, as $d_j \in \{d_i\} \cup \{d_c \sim P_n(w) | c = 1, \ldots, C\}$. $P_n(w)$ is the noise distribution and we set it as unigram distribution raised to 3/4th power, which is the same as Word2Vec. $\mathbb{1}(x)$ is an indicator function defined as:

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if } x \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The time complexity of updating $\mathcal{L}$ is $O(T \log T + M \log M + TC)$ when using hierarchical softmax method for $\mathcal{L}_1$ and $\mathcal{L}_2$, or $O((2T + M)C)$ when using negative sampling method. Based on the complexity analysis, our cascade model is efficient enough and can be applied to MOOC datasets.

4 Experiment

Data Sets We collect the sequential data of two MOOCs from Coursera[1] and China University MOOC[2] respectively. The former is an interdiscipline course called *People and Network*, and the second is called *Introduction to MOOC*. From both courses, we collect subtitles of video clips, forum contents and learners' log of clickstream. Table 1 shows the statistical information of the two MOOCs.

For evaluation, we invite the teaching assistants (TAs) of respective courses to label test samples in advance. Note that our model is unsupervised. Therefore, labeled data (thread-subtitle matching pairs) are only used for evaluation, and we do not utilize dev dataset.

Experimental Setting We compare our embeddings with unsupervised rivals and the labels are only used for evaluation. To ensure fair comparison, we represent documents with their averaged word embeddings. Note that in the training phase, we represent each thread/subtitle with a vector, in order to make the words within a document clustered and close to each other. We evaluate the following methods.

- Bag-of-words(BOW): the classical text representation.
- Word2Vec: word embeddings which leverages word-level coherence and we adopt the CBOW architecture.
- Para2Vec: paragraph embeddings which considers document-level context information. We also adopt CBOW framework.
- Hierarchical Document Vector(HDV): the latest word embeddings with a hierarchical architecture for modeling streaming documents and their contents.
- Cascade Document Representation (CDR): our proposed model which captures both the latent semantics and latent similarity.

We use the hype-parameters recommended by previous literatures. For all the evaluated baselines, we use the same parameter setting. Thus it is fair to make comparison. The window size set in all baselines is 5 by default. The number of negative samples is empirically set as 5. The size of hidden layer is set as 100 for all the methods. We utilize the Precision@K (denoted by P@K) as metric. If the retrieved top-K subtitles hit at least one ground-truth label, we regard it as true; otherwise, it is false. In our experiments, we run 10 times and report the average result for each case.

[1] https://www.coursera.org, which is an educational technology company that offers MOOCs worldwide.

[2] http://www.icourse163.org, which is a leading MOOCs platform in China.

Course Name	#active users	#video clips	#threads	#posts	#words	#clicks
People and Network	10,807	60	219	1,206	121,142	31,096
Introduction to MOOC	3,949	19	557	7,177	480,495	45,642

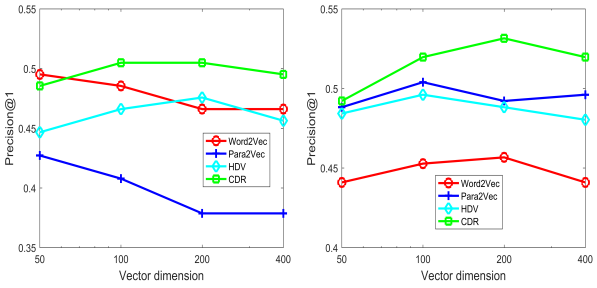
Table 1: Statistics of two MOOC datasets.

Model	*People and Network*			*Introduction to MOOC*		
	P@1	P@3	P@5	P@1	P@3	P@5
BOW	0.437	**0.718**	0.806	0.449	0.811	0.909
Word2Vec	0.485	0.699	**0.816**	0.453	0.826	0.890
Para2Vec	0.408	0.612	0.728	0.504	0.823	0.894
HDV	0.466	0.621	0.777	0.496	0.819	0.913
CDR	**0.505**	0.689	0.786	**0.520**	**0.854**	**0.941**

Table 2: Result of thread-subtitle matching.

Result Firstly we use all the data to learn word embeddings by models. Then the learned word vectors are utilized to calculate the similarity between threads and subtitles, and rank the subtitles. Table 2 reports the results of thread-subtitle matching. We can notice that there are some anomalies in P@3 and P@5 results. It may be for the reason of dataset. In the first MOOC (people and network), video subtitles contain relatively less words, and therefore it is hard to get effective representations. Overall, the proposed models can achieve better performance than baselines, and we highlight the Precision@1 results. Compared to HDV which also considers the streaming documents, our model is better at every task. This indicates our model can effectively capture the latent similarity.

We investigate the effect of number of dimensions, i.e., the size of the neural network's hidden layer. From Fig.2, we find that CDR can achieve better performance than baselines with various numbers of dimensions. In addition, the optimal results can be obtained when the dimension is set as 100 or 200 in both datasets.



(a) People and Network (b) Introduction to MOOC

Figure 2: P@1 of different dimensions.

5 Conclusion

In this paper, we propose an approach to solve a significant problem: how to learn distinguishable representations from word sequences in documents and clickstreams of learners. To model the heterogeneous data, we develop a cascade model which can jointly learn the latent semantics and latent similarity without labeling data. We conduct experiments on two real datasets, and the results demonstrate the effectiveness of our model. Moreover, our model is not limited to MOOC data. For instance, we can adopt the proposed algorithm to streaming documents, e.g. webpage click streams, since our method can model the document-document sequences. We leave this as the future work.

Acknowledgments

This research is supported by the National Research Foundation, Prime Ministers Office, Singapore under its IDM Futures Funding Initiative, China NSFC with Grant No.61532001 and No.61472013, and China MOE-RCOE with Grant No.2016ZD201. Xiaoming Li is the corresponding author. We thank the anonymous reviewers for their insightful comments.

References

Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips. In *EDM*, pages 297–304.

Ashton Anderson, Daniel P. Huttenlocher, Jon M. Kleinberg, and Jure Leskovec. 2014. Engaging with massive online courses. In *WWW*, pages 687–698.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. 2014. Predicting instructors intervention in mooc forums. In *ACL*, pages 1501–1511.

Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic, and Narayan Bhamidipati. 2015. Hierarchical neural language models for joint representation of streaming documents and their content. In *WWW*, pages 248–255.

Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. 2017. Poi2vec: Geographical latent representation for predicting future visitors. In *AAAI*, pages 102–108.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *KDD*, pages 855–864.

Jonathan Huang, Anirban Dasgupta, Arpita Ghosh, Jane Manning, and Marc Sanders. 2014. Superposter behavior in mooc forums. In *L@S*, pages 117–126.

Maximilian Jenders, Ralf Krestel, and Felix Naumann. 2016. Which answer is best?: Predicting accepted answers in mooc forums. In *WWW*, pages 679–684.

Zhuoxuan Jiang, Shanshan Feng, Weizheng Chen, Guangtao Wang, and Xiaoming Li. 2017. Unsupervised embedding for latent similarity by modeling heterogeneous mooc data. In *PAKDD*, pages 683–695.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196.

Tomas Mikolov, Martin Karafit, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Jiezhong Qiu, Jie Tang, Tracy Xiao Liu, Jie Gong, Chenhui Zhang, Qian Zhang, and Yufei Xue. 2016. Modeling and predicting learning behavior in moocs. In *WSDM*, pages 93–102.

Arti Ramesh, Shachi H. Kumar, James R. Foulds, and Lise Getoor. 2015. Weakly supervised models of aspect-sentiment for online course discussion forums. In *ACL&IJCNLP*, pages 74–83.

Lorenzo A. Rossi and Omprakash Gnawali. 2014. Language independent analysis and classification of discussion threads in coursera mooc forums. In *IRI*, pages 654–661.

Gerard Salton and Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *KDD*, pages 1165–1174.

Alyssa Friend Wise, Yi Cui, and Jovita Vytasek. 2016. Bringing order to chaos in mooc discussion forums with content-related thread identification. In *LAK*, pages 188–197.