LTWNN: A Novel Approach using Sentence Embeddings for Extracting Diverse Concepts in MOOCs

Zhijie Wu[1] , Jia Zhu[2], Shi Xu[1], Zhiwen Yan[1], and Wanying Liang[1]
1.School of Computer Science of South China Normal University, Guangzhou 510631, Guangdong, China
2. Zhejiang Normal University , Zhejiang, China,

ABSTRACT

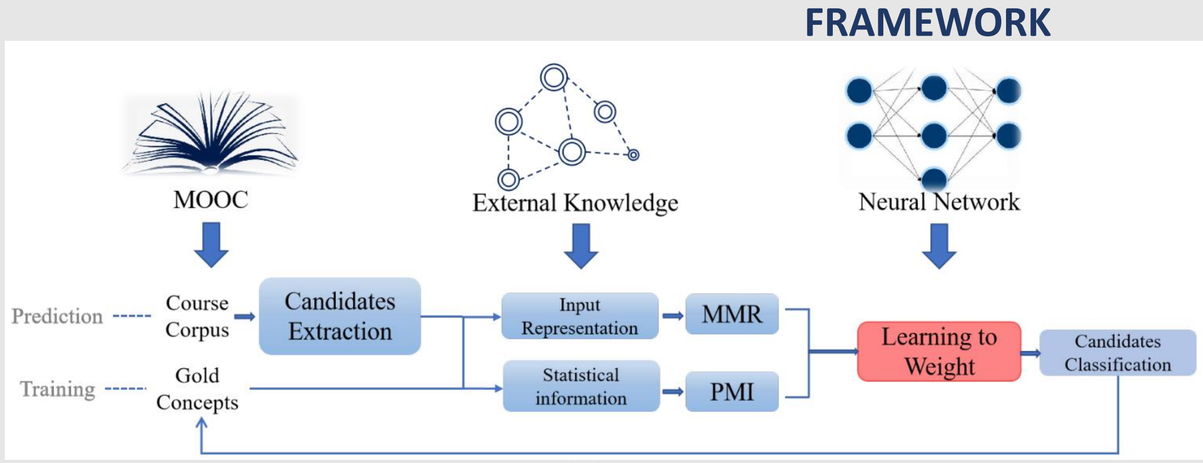
As a global online education platform, Massive Open Online Courses (MOOCs) provide high-quality learning content. It is a challenging issue to design a key course concept for students with different backgrounds. Even though much work concerned with course concept extraction in MOOC has been done, those related works simply utilize external knowledge to get the relatedness of two different candidate concepts. Furthermore, they require the input to belong to multi-document and severely rely on seed sets, in which their model shows poor performance when input is a single document. Addressing these drawbacks, we tackle concept extraction from a single document using **LTWNN**, a novel method **L**earning **t**o **W**eight with **N**eural **N**etwork for Course Concept Extraction in MOOCs. With LTWNN, we make full use of external knowledge via making relatedness between each candidate concept and document by introducing an embedding-based maximal marginal relevance(MMR), which explicitly increases diversity among selected concepts. Moreover, we combine the inner statistical information and external knowledge, in which the neural network automatically learns to allocate weight for them. Experiments on different course corpus show that our method outperforms alternative methods.

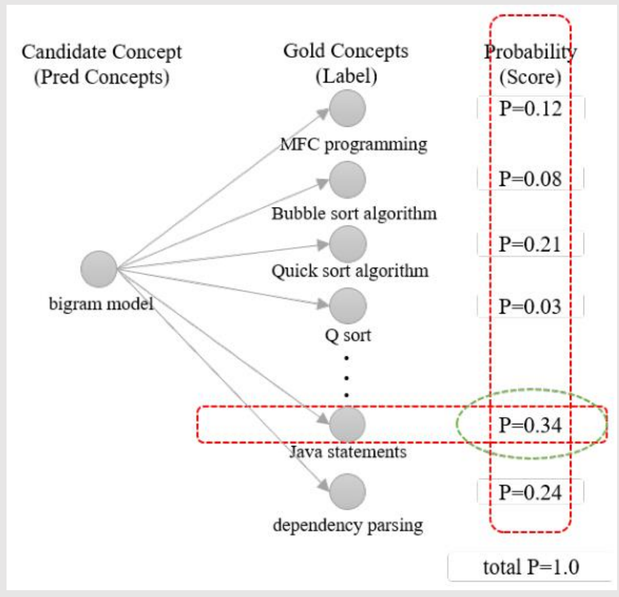
INTRODUCTION

Understanding the overall concept makes it easier to learn the subject and assist in understanding the text for learners. Course concept extraction is non-trivial and challenging due to three reasons, including the single short context problem, the low-frequency problem, and the poor diversity of concepts.

To address the above problems, we propose learning to weight using sentence embedding with neural networks for course concept extraction in this work. The critical aspect of our idea is that it cannot only improve the diversity of extracted course concepts by introducing external knowledge but also automatically learn to weight to leverage inner statistical information and external expertise. First, we extract some keyphrases as candidates by the Part-of-speech (POS) rule template, and we introduce external knowledge to represent each document by sentence embedding model. Then, to improve the diversity of extracted concepts, we introduce the MMR algorithm and change the formula to fit our task. Next, we combine with the score of MMR and statistical information (i.e., PMI), and then our model learns to weight by neural network classifier (e.g., MLP). Finally, in the prediction phase, the MMR score and PMI score of each candidate concept will be the input of the trained model.

FRAMEWORK





(1) PMI

$$PMI(c_1, c_2) = \frac{2 \times freq(c_1, c_2)}{freq(c_1) + freq(c_2)}$$

(2) MMR

$$\mathbf{MMR} := \arg\max_{D_i \in R \setminus S} [\lambda \cdot Sim_1(D_i, Q) - (1-\lambda) \max_{D_j \in S} Sim_2(D_i, D_j)]$$

$$\mathbf{MMR} := \arg\max_{C_i \in C \setminus K} [\gamma \cdot \widehat{\cos}_{sim}(D_i, doc) - (1-\gamma) \max_{C_j \in K} \widehat{\cos}_{sim}(D_i, D_j)]$$

(3) Learning to Weight

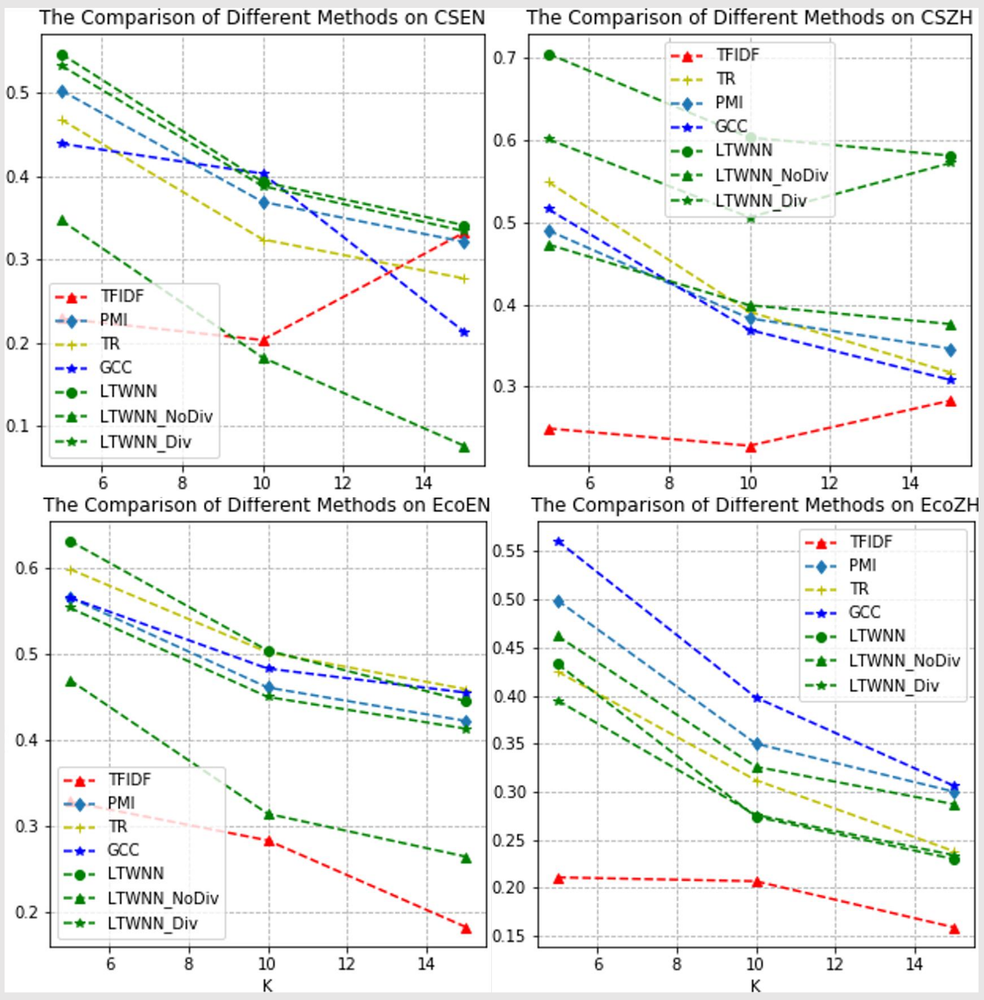
$$\widehat{\cos}(C_i, doc) = 0.5 + \frac{n\cos_{sim}(C_i, doc) - n\cos_{sim}(C, doc)}{\sigma(n\cos_{sim}(C, doc))}$$

$$n\cos_{sim}(C, doc) = \frac{\cos_{sim}(C_i, doc) - \min_{C_j \in C} \cos_{sim}(C_j, doc)}{\max_{C_j \in C} \cos_{sim}(C_j, doc)}$$

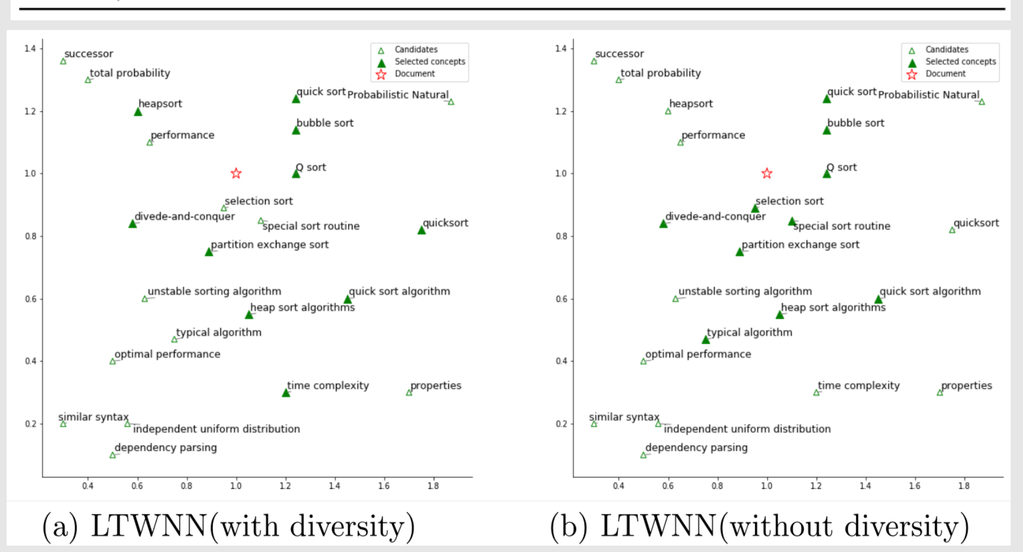
$$p(y_c \mid c) = MLP(c) \qquad pro = softmax(ReLU(XW_h + b_h)) \qquad score = max(pro)$$

RESULTS

For the performance on English data, LTWNN outperforms other methods at the K = 5. Moreover, when the K={10, 15}, LTWNN shows similar performance with the state-of-the-art model. For the performance on dataset CSZH, LTWNN shows apparent robustness and effectiveness over other methods. From the information described in Table, we know the average number of concepts per document is only 1.86, which indicates that the phenomenon of low-frequency and poverty of diversity on the dataset is more obvious than others. Thus, the experiment suggests that LTWNN is effective in solving the problem of low-frequency and poverty of diversity on a single document.



K	Method	CSZH		CSEN		EcoZH		EcoEN	
		MAP	R_p	MAP	R_p	MAP	R_p	MAP	R_p
5	CGP	0.516	0.238	0.509	0.249	**0.561**	**0.297**	0.566	**0.335**
	LTWNN	**0.705**	**0.409**	**0.547**	**0.256**	0.433	0.235	**0.632**	0.295
	W/O PMI	0.601	**0.413**	0.533	0.256	0.395	0.232	0.554	0.298
	W/O MMR	0.473	0.358	0.347	0.135	0.462	0.251	0.470	0.243
10	CGP	0.369	0.135	0.393	**0.276**	**0.398**	**0.181**	0.483	**0.254**
	LTWNN	**0.603**	**0.337**	**0.403**	0.181	0.274	0.162	**0.504**	0.216
	W/O PMI	0.506	0.336	0.388	0.181	0.276	0.158	0.450	0.223
	W/O MMR	0.399	0.311	0.182	0.140	0.326	0.159	0.314	0.191
15	CGP	0.308	0.102	0.213	0.135	**0.306**	**0.140**	**0.455**	**0.224**
	LTWNN	**0.581**	**0.317**	**0.341**	**0.159**	0.230	0.134	0.445	0.202
	W/O PMI	0.572	0.316	0.334	0.158	0.234	0.136	0.413	0.202
	W/O MMR	0.376	0.299	0.076	0.109	0.287	0.133	0.264	0.180



(a) LTWNN(with diversity) (b) LTWNN(without diversity)

CONCLUSIONS

The study is aimed at extracting core knowledge for different background students. The content from MOOC courses is usually rich and complex, which is difficult for students to understand and analyze the knowledge from a global perspective. Course-related concepts represent the core knowledge, which will help students grasp the core knowledge. Moreover, constructing educational knowledge graph based on the course concept entity is helpful for students and teachers, including makes personal education and deep knowledge tracking. In future work, incorporating other external knowledge such as topic knowledge that classifies course knowledge into several groups is an available method to further improve the performance of course concept extraction.