A Multi-Granularity Semantic-Enhanced Model for Concept Extraction on Chinese MOOCs

Anonymous ACL submission

Abstract

As online education becomes popular, open course platforms represented by MOOCs have collected a large number of course videos. How to identify and extract course concepts in MOOC videos accurately has become a fundamental problem in course content analysis and recommendation. However, since the course concepts in video subtitles are complex and diverse, using character features is not enough to understand concept semantics and identify their boundaries. Thus, we propose a Multi-Granularity Semantic-Enhanced (MGSE) model, which unifies information at word and context granularity, to enhance character representations encoded by a pre-trained language model. For word granularity, we design a word assignment policy and a word quality evaluation strategy. For context granularity, we devise a dual-channel attention module to fuse global and similar context information relevant to course concepts. Experimental results on computer courses and economic courses in MoocData show that MGSE outperforms the baselines significantly. The ablation experiment proves that the semantics with various kinds of granularity help the course concept extraction.

1 Introduction

With the development of MOOCs, online education has become an important supplement to classroom education, attracting hundreds of millions of learners. Teaching video is an important component in MOOCs, where the lecture content often starts from a single course concept, and then steps forward to a large number of course concepts. Course concepts are the core elements of the course content. Thus extracting the course concepts from MOOCs video subtitles helps to refine the key information of the videos, which is the fundamental part of the course content analysis and recommendation.

Table 1: POS of words in different entities and course concepts.

Entity	Part of speech
Person	张三 (ZhangSan) /np; 小明 (XiaoMing) /np
Location	北京 (Beijing) /ns; 江西 (Jiangxi) /ns
Organization	教育部 (Ministry of Education) /ni; 中国 (China) /ns; 港口协会 (Port Association) /ni; 财政部 (Ministry of Finance) /ni; 美国 (American) /ns; 卫生组织 (Health Organization) /ni
Concept	绝对 (absolute) /a; 地址 (address) /n; 文件(file) /n; 描述符 (descriptor) /n; 异常 (exception) /a; 自 (auto) /p; 编码器 (encoder) /n; 堆 (heap) /q; buffer /e; main /e; 函数 (function) /n

Term extraction and entity extraction methods based on deep learning are fruitful. These methods encode characters or words at sentence granularity and use CRF (Conditional Random Field) to find the optimal path after label prediction (Huang et al., 2015). However, course concept extraction from video subtitles on Chinese MOOCs has its particularities as follows.

Firstly, the words in Chinese Moocs are domain-specialized, their part of speech (POS) are diverse and there are underlying patterns between the POS (as shown in Table 1). Besides, the Chinese course concept often appears in the form of a phrase. Thus, domain specialization, the pattern rule of the POS, and the tendency to form a phrase are important for candidate word selection in Chinese course concept extraction.

Secondly, Chinese text has no space separator between words as in English. This makes boundary recognition more important for course concept extraction on Chinese Moocs. For instance, when extracting the course concept “自编码器(auto-encoder)”, some course concepts such as “自编

Table 2: The contexts related to course concepts “过程调用(procedure call)”.

那么，再讲讲过程调用。我们说过c语言，可以看成是过程、套过程的一种语言了，在里面反复的做调用，那么可以利用栈并行的这个规律来支持过程调用与返回。实际上这个很简单，大家想想看，过程调用一级，套用一级。过程调用，一般来说，先被调用的过程肯定是后返回，后被调用的过程肯定是先返回，所以它的工作属性跟栈的工作原理很像。所以这样子话呢，我们就可以利用栈来支持过程调用。(Next, let’s talk about **procedure call**. We have mentioned that programming language C can be seen as a language with **procedures** and nested **procedures**. When executing **call to procedure**, the stack mechanism could be used to support **procedure call** and return. Actually, this is very simple. Let’s think about it, when a **procedure** is defined, the **procedure** is nested in its upper **procedure**. if a **procedure** is **called**, a stack is used to save the state of the upper **calling procedure**, pass parameters to the **called procedure**, and store local variables for the currently executing **procedure**.)

码(auto-encoding)”, “编码(encoding)” and “编码器(encoder)” may increase the difficulty of concept recognition. To handle this problem, most approaches introduce word information into the model based on character granularity (Zhang and Yang, 2018; Ma et al., 2019) but they all ignore the different effects of these words.

Lastly, the related course concepts in Chinese MOOCs are dispersed in the whole video subtitle. For example in Table 2, the course concepts “过程调用 (procedure call)”, “过程 (procedure)”, and “调用 (call)” are repeatedly mentioned under relevant contexts. However, few existing works (Xu et al., 2018) consider the relevant context in term or entity extraction tasks.

Moreover, illegal label sequence is a big challenge because Chinese course concepts often consist of many characters and the nested concepts occur frequently in MOOC videos. Considering the particularities mentioned above, we propose a Multi-Granularity Semantic-Enhanced (MGSE) model for concept extraction on Chinese MOOCs. The contributions of this work are as follows.

- We propose MGSE, which unifies semantics on word and context granularity to enhance character representations encoded by the pre-trained language model. Besides, we use masked CRF to alleviate the illegal label sequence.
- For word granularity enhancement, we propose a new word quality evaluation strategy and a novel word assignment strategy. For context granularity, we design a dual-channel attention module to utilize the information relevant to course concepts in both global context and similar context.
- The experimental results on computer courses and economic courses in MoocData show that the MGSE model achieves $F1$ values of 91.05% and 89.34% respectively, outperforming advanced SoftLexicon and FLAT models.

2 Related Work

The course concept extraction task is usually accomplished using the Named Entity Recognition (NER) method. Early NER methods are mainly based on rules and statistics (Stanković et al., 2016; Khan et al., 2016; Pan et al., 2017). Afterward, deep learning methods have a significant advantage for NER task (Kucza et al., 2018; Huang et al., 2021). In this section, we describe the existing work in general and specific domains respectively.

In the general domain, to recognize the entity boundary, most researchers introduced word granularity enhancement methods based on sequence labeling at character granularity. Zhang and Yang (2018) proposed the Lattice LSTM model to enhance the entity boundary information by incorporating words from external lexicons. However, this model only considers words ending with the current character. SoftLexicon model proposed by Ma et al. (2019) differs the word clusters with character position in the word, which ignores word differences in the same word clusters. Moreover, the FLAT model proposed by Li et al. (2020) also encodes word positions in sentences. Based on the pre-training idea of the FLAT model, Lai et al. (2021) proposed Lttice-BERT to improve its focus on words.

Extracting course concepts from MOOC video subtitles is the domain-specific entity extraction task. For example, in the domain of bridge inspection, Li et al. (2021) proposed a named entity recognition method based on the Transformer-BiLSTM-CRF model to address domain issues such as character polysemy, contextual location correlation, and orientation sensitivity in entities. In the domain of craft, Jia et al. (2022) proposed a CNN-BiLSTM-CRF neural network model incorporating domain knowledge such as rules and dictionaries at entity regularity. In the domain of product attribute

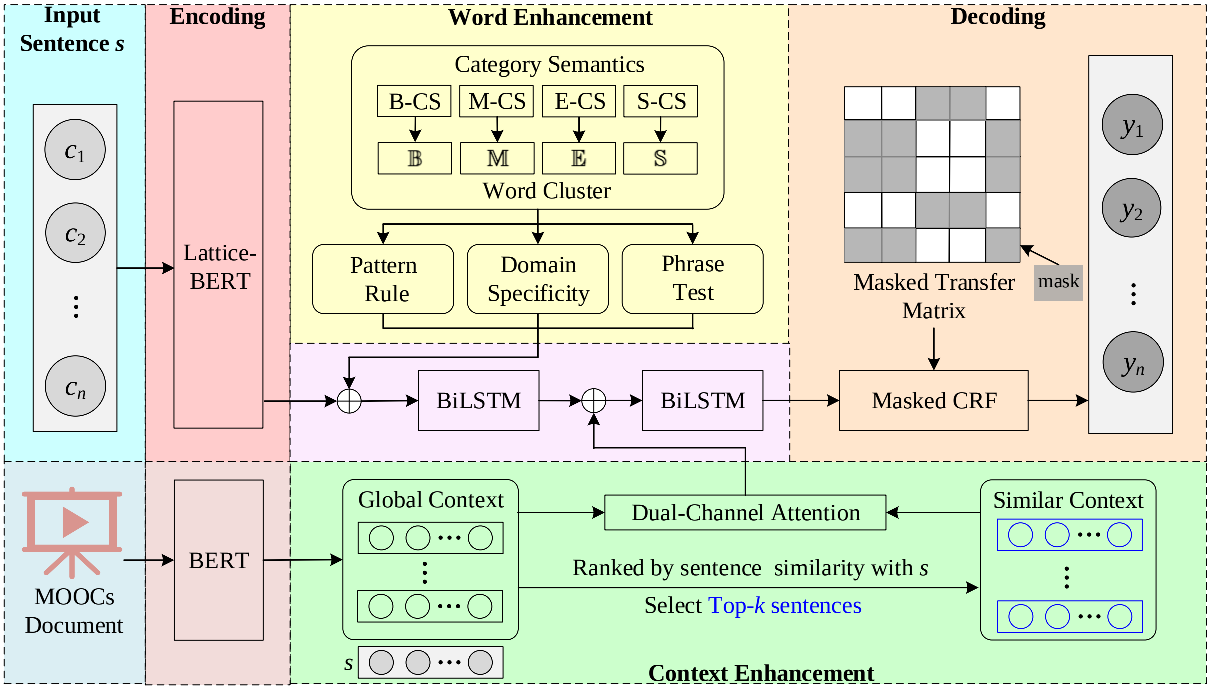

Figure 1: The architecture of MGSE model

extraction, Zhang and Yang (2018) explored the sensitivity of multiple pre-trained language models in terms of text length, attribute value distribution, and noise in domain data.

In the above works, machine learning and deep-learning-based methods mostly focus on the characters in the sentence while labeling the character sequence. Moreover, these models apply either to a general domain or to a specific domain, with less consideration of characteristics of course concepts in video subtitles on MOOCs. Although some models could effectively identify entities or terms by using external resources and knowledge of word granularity, they ignore the effect of different words on course concepts.

3 The MGSE Model

The overall structure of MGSE is shown in Figure 1. Apart from Input, MGSE contains four parts. They are Encoding, Word Enhancement, Context Enhancement, and Decoding.

MGSE uses Lattice-BERT pre-trained language model to encode characters in the input sentence. Moreover, a lexicon is employed when we select candidate words from the input sentence before word enhancement.

In the word enhancement, we design a word assignment strategy to make candidate words separated by character's position. Besides, a word quality evaluation strategy is devised to judge how likely a candidate word is to be treated as a concept.

In the context enhancement, we propose a dual-channel attention mechanism to incorporate global and similar context information of the input sentence. Finally, Masked CRF is employed for decoding.

3.1 Character Encoding

We use Lattice-BERT to enhance the semantics at character granularity. For the input sentence s in the MOOCs video subtitle V, $s = < c_1, c_2, \cdots, c_n >$ and c_i is the i-th character in s. c_i is embedded as a vector representation ${x_i}^{LB} = e_{d_1}^{LB}(c_i)$, where $e_{d_1}^{LB}$ is the mapping table of the character vectors in the Lattice-BERT, and d_1 is the dimension of ${x_i}^{LB}$.

3.2 Word Enhancement

Word enhancement is designed to model the position of the character in a candidate word and to evaluate the likelihood of the word being a course concept or a part of it. It consists of a word assignment unit and a word quality evaluation unit.

3.2.1 Word Assignment

As mentioned before, existing works ignored the different effects of words where the character occurs at different positions. Thus, we assign words

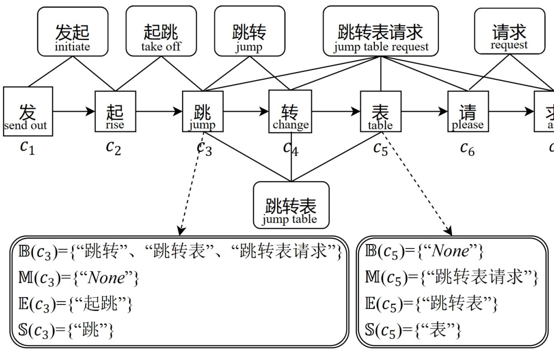

Figure 2: An example of word clusters

Table 3: Category semantics of word clusters

Word Cluster	Category Description
$\mathbb{B}$	Current character occurs at the beginning of these words.
$\mathbb{M}$	Current character occurs at the middle of these words.
$\mathbb{E}$	Current character occurs at the end of these words.
$\mathbb{S}$	Current character is a word.

to different clusters based on the characters' position in words. The steps are as follows. For character c_i in sentence s, we first find candidate words in sentence s by searching the lexicon. Then we assign candidate words to four word clusters $\mathbb{B}(c_i), \mathbb{M}(c_i), \mathbb{E}(c_i)$ and $\mathbb{S}(c_i)$ respectively according to the position of c_i in candidate words. An example is given in Figure 2. The word clusters are defined as follows:

$$\begin{aligned}\mathbb{B}(c_i) =&\{w = [c_i, c_{i+1}, ..., c_l], w \in \mathbb{D}, i < l \leq n\},\\ \mathbb{M}(c_i) =&\{w = [c_j, ..., c_i, ..., c_l], w \in \mathbb{D}\\ &1 \leq j < i < l \leq n\},\\ \mathbb{E}(c_i) =&\{w = [c_j, ..., c_{i-1}, c_i], w \in \mathbb{D}, 1 \leq j < i\},\\ \mathbb{S}(c_i) =&\{w = [c_i], w_i \in \mathbb{D}\},\end{aligned} \tag{1}$$

where $\mathbb{D}$ is a large-scale lexicon. This strategy further enhances the character information and facilitates boundary recognition.

The Category Semantics of Word Clusters. Referring to Ma et al. (2022), we enhance the semantics of word clusters with the prior information as shown in Table 3 to help boundary recognition. There are four categories of word clusters, $\mathbb{B}, \mathbb{M}, \mathbb{E}$, and $\mathbb{S}$, and each category has unique semantics. For example, $\mathbb{B}(c_i)$ is the word cluster in which all words started with the current character c_i. The category semantics of word clusters, denoted by x_l^{CS}, are encoded by BERT.

3.2.2 Word Quality Evaluation

The word quality is used to evaluate the likelihood of words in a word cluster being a course concept or a part of a course concept. The evaluation is carried out from three perspectives based on statistics and rules.

Phrase Measurement. Phrase measurement evaluates the likelihood that a candidate word composed of multiple characters is a complete word, according to the statistics on the MOOCs dataset. In this paper, we evaluate each word w in the word clusters by PMI (Pointwise Mutual Information), which is the co-occurrence frequency of the prefixes and the suffixes making up the word. Specifically, each word $w = \{c_1, c_2, \cdots, c_k\}(k > 1)$ is split into $f_i = c_1, \cdots, c_i$ (prefix) and $b_i = c_{i+1}, \cdots, c_k$ (suffix), where $i = 1, \cdots, k$-1. The phrase score $pm(w)$ of w is defined as follows:

$$pm(w) = \max\{\text{pmi}(f_i, b_i)|i = 1, ..., k\text{-}1\}. \tag{2}$$

Domain Specificity. Domain specificity evaluates the likelihood that a word belongs to a specific domain. Domain-related concepts usually occur with higher frequency in the domain corpus than that in the general corpus. The domain specificity $ds(w)$ of word $w = \{c_1, c_2, ..., c_k\}$ is calculated as follows:

$$ds(w) = \frac{1}{|w|}\sum_{c_i \in w}\log\frac{P^M(c_i)}{P^C(c_i)}, \tag{3}$$

where $|w|$ denotes the number of characters in w, $P^M(c_i)$ and $P^C(c_i)$ denote the probability that the character c_i occurs in the domain corpus M and in the reference corpus C respectively. In this paper, the domain corpus M is the MOOCs dataset, and the reference corpus C is the BCC corpus[1].

Pattern Rule of the POS. Words with different POS have different possibilities to be the whole or part of a course concept. Based on rule-based methods(Pan et al., 2017), we construct a pattern rule to select words for course concepts. Given a sentence s which is split into words with POS, word w in sentence s has a higher possibility of being a course concept or being a part of it if the POS of w satisfies the Parten Rule PR, and the corresponding weight $pr(w)$ is defined as follows:

$$pr(w) = \begin{cases} 1 + \alpha,\ w\ satisfies\ the\ PR \\ 1 - \alpha,\ others \end{cases} \tag{4}$$

[1] http://bcc.blcu.edu.cn/

$$PR = ((((A|N) + |(A|N))|ENG * (NP)? (A|N)*)N)|ENG* \quad (5)$$

where *A*, *N*, *P* and *ENG* denote adjectives, nouns, prepositions, and English characters respectively, and $\alpha \in [0, 1]$ is a predefined parameter.

Comprehensive Quality Assessment. After phrase measurement, domain specificity evaluation, and pattern rule matching for word w, these scores are weighted and summed to calculate the vector representation x_w^W of w as follows: $x_w^W = [W_1 \cdot pm(w) + W_2 \cdot ds(w) + W_3 \cdot pr(w)] \cdot e_{d_2}^W(w)$ where $e_{d_2}^W(w)$ denotes the mapping table from Word2vec, d_2 is the dimension of x_w^W, and W_1, W_2 and W_3 are learnable parameters. The vector representation of a word clusters l, x_l^L , is defined as the mean of the embeddings of all words in l as follows:

$$x_l^L = \frac{4}{Z} \sum_{w \in l} x_w^W, \quad (6)$$

where $Z = \sum_{w \in L}[W_1 \cdot pm(w) + W_2 \cdot ds(w) + W_3 \cdot pr(w)]$ is the normalization factor, $l \subset \{\mathbb{B}(c_i), \mathbb{M}(c_i), \mathbb{E}(c_i), \mathbb{S}(c_i)\}$, and $L = \mathbb{B}(c_i) \cup \mathbb{M}(c_i) \cup \mathbb{E}(c_i) \cup \mathbb{S}(c_i)$.

We concatenated x_l^L with the category semantics x_l^{CS} (the dimension is reduced to the same as x_l^L by a fully connected layer) to obtain the final vector representation x_l^{LCS} of the word cluster l as follows:

$$x_l^{LCS} = [x_l^L; x_l^{CS}]. \quad (7)$$

The lexical representation of characters c_i is the concatenation of all representations on various word clusters as follows:

$$x_i^{SEG} = [x_{\mathbb{B}}^{LCS}; x_{\mathbb{M}}^{LCS}; x_{\mathbb{E}}^{LCS}; x_{\mathbb{S}}^{LCS}]. \quad (8)$$

Finally, the Lattice-BERT vector representation x_i^{LB} and the lexical representation x_i^{SEG} of character c_i are concatenated together to obtain the final representation x_i^C of c_i :

$$x_i^C = [x_i^{LB}; x_i^{SEG}], \quad (9)$$

x_i^C incorporates the information about the candidate words where c_i occurs, which can enhance the semantic expression and the boundary discrimination for the proposed model.

The first layer BiLSTM is used to model the inter-character dependencies in the sentence. The hidden representation h_i^C of character c_i is as follows:

$$h_i^C = \left[\overrightarrow{\mathrm{LSTM}}(x_i^C); \overleftarrow{\mathrm{LSTM}}(x_i^C)\right], \quad (10)$$

where h_i^C considers only the sentence context in which the character occurs.

3.3 Context Enhancement

The entire MOOC document V in which a course concept occurs is helpful for course concept extraction. However, MOOC documents are usually long and the process of the instructor's lecture is relatively free, i.e. adding or switching topics depending on student reception and classroom scenarios, which results in the context related to a certain course concept scattered at different time points in the videos. We design a dual-channel attention mechanism module to model context semantics in Chinese MOOCs.

We rank all sentences in the MOOCs document where the input sentence s occurs based on the F_{BERT} score from BERTScore Zhang et al. (2019), and the top-k sentences that are most semantically relevant to s are selected as the Similar Context S.

Specifically, each sentence in the similar context S is embedded by BERT, and its dimension is reduced to the same as h_i^C, denoted by h_j^B ($j = 1, ..., k$). The attention mechanism is employed to get the semantic of S, denoted by h_i^S, concerning the character c_i.

$$\alpha_{i,j} = \frac{exp(score(h_i^C, h_j^B))}{\sum_{t=1}^{k} exp(score(h_i^C, h_t^B))}$$
$$score(h_i^C, h_j^B) = \frac{(h_j^B)^T \cdot h_i^C}{\sqrt{d_3}} \quad (11)$$
$$h_i^S = \sum_{j=1}^{k} \alpha_{i,j}\, h_j^B$$

where d_3 is the dimension of the sentence vector.

Similarly, we can obtain the global context embedding h_i^G based on all sentences in V. Both of h_i^G and h_i^S are concatenated together to obtain the context vector h_i^{SG} of character c_i :

$$h_i^{SG} = [h_i^S; h_i^G]. \quad (12)$$

Finally, the representation h_i^C from the first layer of BiLSTM and its context vector h_i^{SG} are concatenated, which is fed into the second layer of BiLSTM to obtain the final representation h_i^{CSG} of the character c_i as follows:

$$h_i^{CSG} = [\overrightarrow{\mathrm{LSTM}}([h_i^C; h_i^{SG}]); \overleftarrow{\mathrm{LSTM}}([h_i^C; h_i^{SG}])]. \quad (13)$$

3.4 Masked CRF Decoding

For decoding, the label sequence should satisfy some constraints when extracting course concepts by the sequence labeling method. For example, "B" (the first character in the course concept) is before "M" (the middle character in the course concept), thus the label sequence "O M O" ("O" is a non-course concept character) is an illegal path. Although the CRF model has its constraint for labels, the constraint is relatively weak. To eliminate the illegal transfers in MGSE, instead of using random initialization, we modify the transfer matrix of CRF by a masked matrix, where all illegal transfers are masked by a very small transfer probability. As shown in Figure 1, the transfer probability of all illegal transfers (gray part) in the mask transfer matrix is set to a very small value ϵ.

Let Ω be the set of all illegal transfers, we use equations 14 to obtain the masked transfer matrix $\overline{\mathbf{A}}$ for a given transfer matrix $\mathbf{A}$, where $\epsilon << 0$, and $\delta_{i,j}$ is the trainable transfer probability.

$$\overline{\delta}_{i,j} = \begin{cases} \epsilon, \ if(i,j) \in \ \Omega \\ \delta_{i,j}, \ otherwise \end{cases} \tag{14}$$

For the input sentence $s =< c_1, c_2, ..., c_n >$ and the predicted label sequence $\hat{y} =< y_1, y_2, ..., y_n >$, the scores of $\hat{y}$ is calculated as follows.

$$Score(s, \hat{y}) = \sum_{i=0}^{n} \overline{\delta}_{y_i, y_{i+1}} + \sum_{i=1}^{n} p_{i,y_i} \tag{15}$$

where the $\overline{\delta}_{y_i, y_{i+1}}$ is the probability that label y_i transfers to y_{i+1} in the masked transfer matrix $\overline{\mathbf{A}}$, and p_{i,y_i} is the probability that c_i has label y_i , which is the output of softmax layer with h_i^{CSG} as input. Suppose all possible paths are denoted by Y and all illegal paths are denoted by I, the Masked CRF restricts the "path space" to all legal paths Y/I. The model is trained by maximizing the probability of the ideal path y in Equation 16. The path y^* with the highest probability is calculated by Equation 17 when testing.

$$p(y|s) = \frac{exp(Score(s, y))}{\sum_{\bar{y} \in Y/I} exp(Score(s, \bar{y}))} \tag{16}$$

$$y^* = argmax_{\bar{y} \in Y/I} \ Score(x, \bar{y}) \tag{17}$$

4 Experiments

4.1 Datasets and Evaluation Methods

We use the MoocData, an open course video subtitle datasets[2] provided in Pan et al. (2017). MoocData consists of four sub-datasets, that is the computer science course subset CSZH (in Chinese) and CSEN (in English), and the economics course subset EcoZH (in Chinese) and EcoEN (in English). Each subset contains course video subtitle documents and a collection of manually constructed course concepts. Since our work focuses on Chinese MOOC video subtitles, for comparability, we follow Huang et al. (2021) and select the course numbered "14" in CSZH for model training and evaluation. Moreover, we examine the domain adaptability of the MGSE model on the course subset EcoZH. The dataset was annotated by a remotely supervised method and checked by a group of postgraduates majoring in computer science. The annotated data are divided into train, test, and validation sets according to the ratio of 8: 1: 1. The statistics of the datasets are shown in Table 4. The precision rate P, recall rate R and $F1$ value are chosen as the evaluation methods.

[2]http://moocdata.cn/data/concept-extraction

Table 4: Datasets (CSZH, course No. 14)

Datasets	#Video	#Sentence	#Entity	#Char
Train set	104	4,650	6,804	188,615
Test set	13	580	856	22,975
Validation set	13	580	827	23,093

4.2 Hyper-parameter Settings and Baselines

The hyper-parameters in MGSE are reported in Table 5. The initial learning rate is set to 0.0015 and fine-tuned with model training. α and top-k are set to 0.05 and 10 which depend on model performance on the validation set. ϵ is set to -100 referring to (Wei et al., 2021).

Table 5: Hyper-parameters in the MGSE model

Parameter	Value	Parameter	Value
Initial learning rate	0.0015	Optimizer	Adam
LSTM hidden dim(h_i^C, h_i^{CSG})	200	Dropout	0.5
LSTM layer	2	d_1	768
d_2	50	d_3	200
Dimension of x_l^{LCS}	100	α	0.05
Top-k	10	ϵ	-100

To comprehensively evaluate the model in this paper, the relevant methods on named entity recognition and course concept extraction were selected as the baselines, including BERT-BiLSTM-CRF, Lattice LSTM, LR-CNN, WC-LSTM, CGN, SoftLexicon (LSTM) + BERT, and FLAT. A detailed description of the baselines is presented in the Appendix A.

Table 6: Experimental results

Models	CSZH			EcoZH		
	P	**R**	**F1**	**P**	**R**	**F1**
Lattice LSTM	85.21	88.03	86.60	-	-	-
LR-CNN	85.55	89.81	87.63	-	-	-
CGN	85.10	90.45	87.69	-	-	-
WC-LSTM	86.20	89.57	87.86	-	-	-
BERT-BiLSTM-CRF	85.54	90.40	87.90	91.60	81.63	86.33
SoftLexicon+BERT	85.63	91.11	88.29	91.67	83.54	87.42
FLAT	86.54	90.63	88.53	90.09	84.24	87.07
MGSE	**89.65**	**92.49**	**91.05**	**91.95**	**86.87**	**89.34**

Table 7: Ablation experimental results

Model	P	R	F1
MGSE	**89.65**	**92.49**	**91.05**
- Lattice BERT	87.90	88.75	88.32 (-2.73%)
- Words Quality Evaluation	89.34	90.96	90.14 (-0.91%)
- Context Information	89.05	90.45	89.74 (-1.31%)
- Category Semantics	89.59	92.32	90.93 (-0.12%)
- Masked	89.47	91.66	90.55(-0.50%)

4.3 Experimental Results

The experimental results for each model are shown in Table 6 where MGSE achieves the best results on P, R, and $F1$ values. Each result is an average of 5 independent runs. The result analysis is as follows:

(1) Pre-training model and CRF decoding method are more helpful for course concept extraction. Models like Lattice LSTM, LR-CNN, CGN, and WC-LSTM introduce word information at the character granularity, with $F1$ values of 1.30%, 0.27%, 0.21% and 0.04% lower than that of BERT-BiLSTM-CRF respectively, which indicates that pre-training model BERT and decoding model CRF are important for course concept extraction.

(2) The way of introducing word information has a great influence on concept extraction. Although Lattice LSTM, LR-CNN, WC-LSTM, and CGN are word enhancement models, LR-CNN, WC-LSTM, and CGN are proposed to address the problems of word conflict, the inability of parallel batch training, and the inefficient utilization of word information in Lattice LSTM respectively, with $F1$ values improved by 1.03%, 1.26% and 1.09%, compared with the Lattice LSTM model.

(3) Overall, the FLAT model is outperformed in introducing word information. SoftLexicon (LSTM) + BERT model and FLAT model introduce word information in different ways, however, the former encodes character position in the word, and the latter encodes word position in the sentence. In terms of performance, FLAT has 0.91% higher $F1$ values compared to SoftLexicon (LSTM) + BERT.

(4) Multi-granularity semantic enhancement provides useful information for the semantics and boundaries recognition of course concepts. MGSE model improves the $F1$ value by 2.76% and 2.52% compared to SoftLexicon (LSTM) + BERT and FLAT respectively, indicating that the combination of semantics with multiple granularities at word and context can effectively enhance the semantic representation of course concepts, and locate the boundaries of course concepts more accurately. At word granularity, the importance of words in word clusters is considered comprehensively by word quality evaluation. At context granularity, the similarity context and global context of candidate concepts are introduced into the dual-channel attention mechanism, which helps the model obtain richer semantics and cope with more complex contexts. For decoding, Masked CRF restricts illegal paths better than traditional CRF.

(5) The MGSE model is good at domain adaptability. To verify its adaptability on different course domains, we directly apply the POS pattern rule and hyper-parameters constructed or trained on CSZH to EcoZH. Compared to the BERT-BiLSTM-CRF, SoftLexicon (LSTM) + BERT, and FLAT that performed well, the MGSE model still has 3.01%, 1.92%, and 2.27% higher $F1$ values respectively.

4.4 Ablation Experiments

To verify the role and effect of each module in the MGSE model, ablation experiments are conducted in this section. The model using Bert instead of the pre-trained Lattice-BERT for sentence encoding is denoted as: - Lattice BERT; the model removes word quality evaluation module, dual-channel attention module or semantics of word clusters is denoted as: - Words Quality Evaluation, - Context information or - Category Semantics, respectively; the model replaces Masked CRF with CRF is denoted as: - Masked. The results of the ablation experiment are shown in Table 7.

The results in Table 7 show that the $F1$ of MGSE decreases by 2.73% after removing the Lattice-BERT model, which indicates that the introduction of word-lattice structure enriches the character representation. The $F1$ values decreased by 0.91%, 0.12%, 1.31%, and 0.50% after removing the word quality evaluation module, the category semantic module of word clusters, the dual-channel attention module, and the Masked CRF model, which indicates that the semantic enhancement methods with these modules at different granularities are suitable for course concept extraction from Chinese MOOCs video subtitle.

4.5 Case Analysis

Some extraction cases of the MGSE model are shown in Table 8. In Case 1, the SoftLexicon model annotates "内存访问地址 (Memory Access Address)" as a sequence of "B E M M M E", where the label of character "存(Store)" is identified as "E", resulting in the whole path containing an illegal transfer "E M". Although the nested concept "内存(Memory)" was extracted, it is an incomplete concept in this sentence. Similarly, the FLAT model labels "循环体 (Loop Body)" in Case 2 as "B M O", where "M O" is also an illegal transfer, resulting in incomplete extraction of concept. In both cases, SoftLexicon and FLAT models not only get some illegal transfers but also make some mistakes on the boundary identification. The MGSE model gets the right answers by improving the identification of course concept boundaries through multiple granularity semantics and eliminating the illegal paths by Masked CRF.

Table 8: Cases extracted by the MGSE model

No.	**Results annotated by MGSE model (the labels in brackets are the ideal labels)**
1	因/O(O) 为/O(O) 内**/B(B)** 存**/M(M)** 访**/M(M)** 问**/M(M)** 地**/M(M)** 址**/E(E)** ，/O(O) 它/O(O) 并/O(O) 不/O(O) 存/O(O) 在/O(O) 。/O(O)
2	我/O(O) 在/O(O) 循**/B(B)** 环**/M(M)** 体**/E(E)** 内/O(O) 部/O(O) 做/O(O) 完/O(O) 这/O(O) 个/O(O) 计/O(O) 算/O(O) 。/O(O)
3	压**/B(B)** 栈**/E(E)** 实/O(O) 际/O(O) 上/O(O) 就/O(O) 是/O(O) 一/O(O) 个/O(O) 反**/B(O)** 操**/M(O)** 作**/E(O)** 。/O(O)
4	触/O(O) 发/O(O) 一/O(O) 个/O(O) 读**/B(O)** 地**/M(O)** 址**/E(O)** 不/O(O) 对/O(O) 齐/O(O) 的/O(O) 异**/B(B)** 常**/E(E)** 。/O(O)

In Case 3, the course concept "压栈 (Push into Stack)" is accurately identified by the MGSE model, but "反操作 (Reverse Operation)" is additionally identified as a course concept. Similarly, the MGSE model also extracts "读地址 (Read Address)" in Case 4 as a course concept. As the model encounters course concepts like "并发操作 (Concurrent Operation)" and "内存地址 (Memory Address)" during the training process, "反操作 (Reverse Operation)" and "读地址 (Read Address)" are close to these course concepts in terms of semantics and composition, so that they are mistakenly considered as course concepts. In addition, there is also ambiguity regarding whether "反操作 (Reverse Operation)" and "读地址 (Read Address)" are course concepts or not, which poses a new challenge for course concept extraction models such as MGSE.

5 Conclusion

We propose the MGSE model to meet the characteristics of course concepts in Chinese MOOC video subtitles. The MGSE model improves the semantics expression and boundary recognition for course concepts by introducing semantic information at multi-granularity such as character, word, and context. To discriminate the candidate words where a character occurs at different positions, we propose a word assignment strategy to put them in different word clusters. We design a new word quality evaluation strategy to enhance semantics at word granularity on three aspects such as phrase measurement, domain specificity, and pattern rule of the POS. In addition, we propose a dual-channel attention module, which incorporates global context and similar context, to enhance semantics at context granularity. For decoding, we use masked CRF to eliminate illegal label sequences. The experimental result shows that by combining the semantic information at character, word, and context granularity, the MGSE model outperforms the baselines in extracting course concepts from Chinese MOOC video subtitles.

6 Limitations

The case study on MGSE reveals that some words or phrases are similar to course concepts in terms of semantics and composition, which are difficult to extract for MGSE. In addition, MGSE cannot identify the importance of a concept for the MOOCs document, thus, the extracted concepts cannot represent the core content of the MOOCs video subtitle. Furthermore, the performance of MGSE decreases when it transfers to courses in the domain different from the training courses.

References

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, pages 4982–4988.

Chao Huang, Quanlong Li, Yuanlong Chen, and Dechen Zhan. 2021. An effective method for constructing knowledge graph of online course. In *2021 4th International Conference on Big Data and Education*, pages 12–18.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Meng Jia, Peiyan Wang, Guiping Zhang, and Dongfeng Cai. 2022. Named entity recognition for process text. *Journal of Chinese Information Processing*.

Muhammad Tahir Khan, Yukun Ma, and Jung-jae Kim. 2016. Term ranker: A graph-based re-ranking approach. In *FLAIRS Conference*, pages 310–315.

Maren Kucza, Jan Niehues, Thomas Zenkel, Alex Waibel, and Sebastian Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. In *Interspeech*, pages 2072–2076.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021. Lattice-bert: leveraging multi-granularity representations in chinese pre-trained language models. *arXiv preprint arXiv:2104.07204*.

Ren Li, Tong Li, Jianxi Yang, Tianjin Mo, Shixin Jiang, and Dong Li. 2021. Bridge inspection named entity recognition based on transformer-bilstm-crf. *Journal of Chinese Information Processing*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019. An encoding strategy based word-character lstm for chinese ner. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2379–2389.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label semantics for few shot named entity recognition. *arXiv preprint arXiv:2203.08985*.

Ruotian Ma, Minlong Peng, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *arXiv preprint arXiv:1908.05969*.

Liangming Pan, Xiaochen Wang, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. 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)*, pages 875–884.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 507–514.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3830–3840.

Tianwen Wei, Jianwei Qi, Shenghuan He, and Songtao Sun. 2021. Masked conditional random fields for sequence labeling. *arXiv preprint arXiv:2103.10682*.

Jun Wu, Yao Cheng, Han Hao, Aizezi Ailiyaer, Feixue Liu, and Yipo Su. 2020. Automatic extraction of chinese terminology based on bert embedding and bilstm-crf model. *Journal of the China Society for Scientific and Technical Information*.

Guohai Xu, Chengyu Wang, and Xiaofeng He. 2018. Improving clinical named entity recognition with global neural attention. In *Web and Big Data: Second International Joint Conference, APWeb-WAIM 2018, Macau, China, July 23-25, 2018, Proceedings, Part II 2*, pages 264–279. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

A Appendix

The details of the **baselines** used in this paper are introduced as follows.

(1) BERT-BiLSTM-CRF. The pre-trained BERT language model extracts contextual features of characters, which improves extraction performance effectively. The model is widely used for named entity extraction in various domains. For example, Wu et al. (2020) used this model to extract Chinese professional terms; Huang et al. (2021) applied this model on MOOCs to extract course concepts from video subtitles.

(2) Lattice LSTM. Errors coming from Chinese word separation impair the performance of NER models. To address this issue, Zhang and Yang (2018) proposed a lexical enhancement model, which effectively alleviates this problem by integrating candidate words into the character-based approach with the LSTM network.

(3) LR-CNN. To reduce word conflicts in Lattice LSTM, Gui et al. (2019) used CNN to stack and encode characters, and incorporated lexical information with an attention mechanism.

(4) WC-LSTM. To address parallel batch training in Lattice LSTM, Liu et al. (2019) adopted four strategies to fix the word representation.

(5) CGN. Considering the inefficient use of words in Lattice LSTM, Sui et al. (2019) exploit word knowledge to fuse word information into character representations with a graph attention network GAN, which is based on a collaborative graph network consisting of an encoding layer, a graph network layer, a fusion layer, and a decoding layer.

(6) SoftLexicon (LSTM) + BERT. To reduce information loss in words, Ma et al. (2019) introduced character position in words. In addition, considering the advantages of pre-trained models in character representation, they combined the SoftLexicon (LSTM) model with BERT, naming as "SoftLexicon (LSTM) + BERT".

(7) FLAT. Li et al. (2020) used the word-lattice structure to integrate word-level information into the character-level and encoded the relative position of words in sentences.