# Distantly Supervised Course Concept Extraction in MOOCs with Academic Discipline

**Mengying Lu**[1], **Yuquan Wang**[2], **Jifan Yu**[2]*, **Yexing Du**[3], **Lei Hou**[2], **Juanzi Li**[2]
[1]SIGS, Tsinghua Univerisity, Shenzhen 518055, China
[2]DCST, Tsinghua Univerisity, Beijing 100084, China
[3]DCST, Beijing University of Chemical Technology, Beijing 100029,China
`{lumy22, yujf21}@mails.tsinghua.edu.cn` `yuq406@gmail.com`
`{houlei, lijuanzi}@tsinghua.edu.cn` `duyexing@buct.edu.cn`

## Abstract

With the rapid growth of Massive Open Online Courses (MOOCs), it is expensive and time-consuming to extract high-quality knowledgeable concepts taught in the course by human effort to help learners grasp the essence of the course. In this paper, we propose to automatically extract course concepts using distant supervision to eliminate the heavy work of human annotations, which generates labels by matching them with an easily accessed dictionary. However, this matching process suffers from severe noisy and incomplete annotations because of the limited dictionary and diverse MOOCs. To tackle these challenges, we present a novel three-stage framework DS-MOCE, which leverages the power of pre-trained language models explicitly and implicitly and employs discipline-embedding models with a self-train strategy based on label generation refinement across different domains. We also provide an expert-labeled dataset spanning 20 academic disciplines. Experimental results demonstrate the superiority of DS-MOCE over the state-of-the-art distantly supervised methods (with 7% absolute F1 score improvement). Code and data are now available at `https://github.com/THU-KEG/MOOC-NER`.

## 1 Introduction

Course concept extraction in Massive Open Online Courses (MOOCs) aims to recognize high-quality knowledge concepts and subject terms taught in the course. Automatically extracting course concepts can help students better understand knowledgeable concepts of the course and reduce the burden of teacher workloads (Butt and Lance, 2005). It is a core task in course content analysis and MOOC knowledge graph construction, which is a fundamental step to building AI-driven MOOC systems with various downstream applications such as course recommendation and question answering (Song et al., 2021). However, MOOCs' explosive growth, like the number of online courses, which grew from 13.5k in 2019 to 19.4k in 2021[1], makes it expensive and tedious to annotate course corpus manually. Therefore, there is a clear need to achieve automatically consistent and accurate course concept extraction in MOOCs to eliminate the heavy work of human annotations.

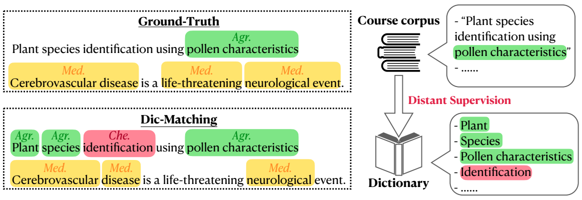

Ground-Truth
Agr.
Plant species identification using pollen characteristics
Med.
Med.
Med.
Cerebrovascular disease is a life-threatening neurological event.
Dic-Matching
Agr.
Agr.
Che.
Agr.
Plant species identification using pollen characteristics
Med.
Med.
Med.
Cerebrovascular disease is a life-threatening neurological event.
Course corpus
- "Plant species identification using pollen characteristics"
- ......
Distant Supervision
Dictionary
- Plant
- Species
- Pollen characteristics
- Identification
- ......

Figure 1: An example of distant labels obtained with a dictionary, suffering from noisy and incomplete annotations. *Che.* corresponds to *Chemistry* with red color, *Agr.* to *Agriculture* with green, and *Med.* to *Medicine* with yellow.

Early works for course concept extraction in MOOCs include graph propagation (Pan et al., 2017; Lu et al., 2019) and statistical ranking methods (Wu et al., 2022; Albahr et al., 2021). Recently, distant supervision has been proposed for the automatic generation of training labels. As shown in Figure 1, the labeling procedure matches the tokens in the course corpus with concepts in an easily accessed dictionary. However, this matching process suffers from two challenges: (1) **noisy annotation** where a mention can be low-quality (i.e., the mention of 'plant' and 'species' of the first instance) or unrelated to the field of the course (i.e., the mention of 'identification' from *Chemistry* but this instance is about *Agriculture*); and (2) **incomplete annotation** where a mention can be matched partially (i.e., the mention of 'cerebrovascular disease' and 'neurological event' of the second instance) or missed completely (i.e., the mention of 'life-threatening' )

*Corresponding author.

[1] `https://www.classcentral.com/report/moocs-stats-and-trends-2021/`

*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 13044–13059
July 9-14, 2023 ©2023 Association for Computational Linguistics

| Dataset | Types | F-1 | P | R |
|---|---|---|---|---|
| CoNLL03 | 4 | 59.61 | 71.91 | 50.90 |
| Tweet | 10 | 35.83 | 40.34 | 32.22 |
| BC5CDR | 2 | 71.98 | 93.93 | 58.35 |
| MOOCs | 20 | 16.85 | 12.50 | 25.84 |

Table 1: Pioneer experiments for distantly supervised performance (Liang et al., 2020) on CoNLL03 (Tjong Kim Sang, 2002), Tweet (Strauss et al., 2016), BC5CDR (Shang et al., 2018), and MOOCs.

due to the limited coverage of dictionary.

Several training paradigms have been employed in Distantly Supervised NER (DS-NER), such as reinforcement learning (Yang et al., 2018) and bagging active learning (Lee et al., 2016) to address the noise annotation; concept expansion (Yu et al., 2019, 2020b; Wang et al., 2019) and positive-unlabeled learning (Peng et al., 2019; Zhou et al., 2022) to address the incomplete challenge. Unfortunately, the previous studies assume a high precision and reasonable recall after distantly supervised label generation. However, severe low-precision and low-recall are reported in MOOCs according to pioneer experiments and comparison with other benchmarks in Table 1. It indicates that there are more noise and incomplete annotations in MOOCs, which significantly hurt following model training performance, thus making the advanced DS-NER approaches fail to cope with the two challenges.

Our analysis yields that the limited dictionary and diverse MOOCs lead to more noise and incomplete annotations[2]. First, the dictionary lacks sufficiently extensive coverage because of MOOCs' rapid growth and missing criteria. Therefore, the out-of-dictionary, low-quality concepts will consequently render more course concepts unmatched and false-positive noisy annotations during matching. Second, MOOCs can span 20 or even more academic disciplines (Mohd Salamon et al., 2016), producing unrelated noisy annotations across different open domains. Additionally, the uneven concept distribution and semantic differences among varied disciplines are different, imposing significant challenges to training an effective and accurate model.

To address the two challenges, we propose a novel three-stage framework DS-MOCE to distantly supervised extract course concepts in MOOCs across different domains. Our framework consists of (1) **Discipline-aware Dictionary Empowerment** which employs prompt-based learning to explicitly generate concept distribution over diverse MOOC domains and implicitly enhance the dictionary's limited capability; (2) **Distant Supervision Refinement** which removes unrelated noise with much higher precision annotations for model training; and (3) **Discipline-embedding Models with Self-training** to deal with noise iteratively while finding incomplete mentions based on semantic knowledge and syntactic information of pre-trained language models (PLMs) and positive-unlabeled learning (PUL).

For evaluation, we provide an expert-labeled dataset spanning 20 academic disciplines, which contains 522 expert-annotated sentences from 17 courses with $15,375$ course concepts.

Our contributions include 1) a novel three-stage framework to distantly supervised extract course concepts in MOOCs across different domains to eliminate the heavy work of human annotations; 2) a distant supervision refinement method to discard unrelated field noise and discipline-embedding models with a self-training strategy to remove noise iteratively and address the incomplete challenge based on PUL; 3) an expert-labeled dataset with the excellent performance of our DS-MOCE framework over existing distantly supervised methods, with one implementation report of 7% absolute F1 score improvement.

## 2 Problem Formulation

Following Pan et al. (2017), we give some necessary definitions and then formulate the problem of distantly supervised course concept extraction.

A **course corpus** is composed of $n$ courses from different academic disciplines, denoted as $D = \{C_i\}_{i=1}^{n}$, where $C_i$ is one course. Each course $C_i = \{S_i, F_i\}$ consists of two parts, where $F_i = [f_{i1}, \dots, f_{ik_i}]$ is course related academic disciplines, and $S_i = \{v_{ij}\}_{j=1,\dots,n_i}$ is composed of $n_i$ course video subtitles, where $v_{ij}$ stands for the $j$-th video subtitles. Finally, we get all academic fields $F = \{f_i\}_{i=1,\dots,k}$ related to course corpus $D$, so $k$ is the number of academic disciplines.

A **dictionary** $T = \{T_i\}_{i=1,\dots,k}$, where $T_i = \{c_{ij}\}_{j=1,\dots,m_i}$ is composed of $m_i$ course concepts $c_{ij}$ in academic disciplines $f_i$.

**Distantly Supervised Course Concept Extraction in MOOCs** is formally defined as follows. Given the course corpus $D$ and dictionary $T$, for

[2] For more case explanations, see Appendix A.1

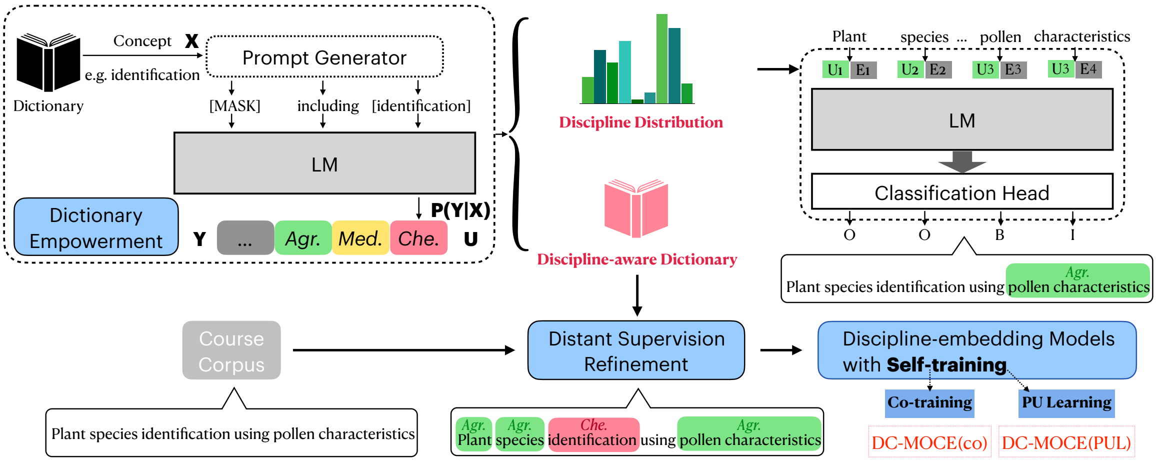


Figure 2: Our proposed three-stage framework **DS-MOCE** for distantly supervised course concept extraction in MOOCs, which includes (1) **Discipline-aware Dictionary Empowerment**; (2) **Distant Supervision Refinement** and (3) **Discipline-embedding Models with Self-training**. For two advanced implementations: **DS-MOCE (co)** means our framework with two student-teacher networks co-training self-train strategy, and **DS-MOCE (PUL)** means our framework adding positive-unlabeled learning loss self-train strategy.

each course $C_i$ in $D$, the objective is to extract $F_i$ discipline-related and high-quality course concepts from video subtitles $S_i$.

## 3 The DS-MOCE Framework

Considering the limited dictionary and diverse MOOCs, it is natural not to ignore the academic discipline characteristics for distantly supervised course concept extraction in MOOCs. As shown in Figure 2, we propose a three-stage framework DS-MOCE, which includes 1) **Discipline-aware Dictionary Empowerment** to transfer the power of PLMs to the dictionary; 2) **Distant Supervision Refinement** which considers academic disciplines to tackle the unrelated field noise explicitly; and 3) **Discipline-embedding Models** to fully exploit the power of PLMs with concept distribution to implicitly handle the noise and incomplete challenges, which then can be integrated with two advanced DS-NER implementations. One employs a co-training strategy to deal with the noise iteratively, denoted as **DS-MOCE(co)**. The other employs PUL to deal with the incomplete problem, denoted as **DS-MOCE(PUL)**.

### 3.1 Discipline-aware Dictionary Empowerment

Before distant supervision, we design a preceding step to conduct discipline classification for each concept in the dictionary with prompt-based learning (Liu et al., 2021b), hoping to transfer semantic knowledge from the language model (LM) to the dictionary. Formally, taking the input of each concept $c_i$ in the dictionary $T = \{c_i\}_{i=1,...,m}$, the classification returns a ranked list of related disciplines $F_{c_i} \subset F$ and outputs $p_j$ for $f_j \in F = \{f_j\}_{j=1,...,k}$ to indicate its likelihood to be related to $f_j$ discipline:

$$p_j(x^{'}) = LM(f_{fill}(x^{'}, f_j); \theta) \qquad (1)$$

where $x^{'} = f_{prompt}(c_i)$ is a prompt with the concept $c_i$ filled template slot $[concept]$, and function $f_{fill}(x^{'}, f_j)$ fills in the slot $[MASK]$ with the potential answer $f_j$. For example (Figure 2), in one case of discipline classification where $c_i$ =“identification”, the template is designed as “$[MASK]$ including $[concept]$”. Then $x^{'}$ would become “ $[MASK]$ including identification”, and we calculate the probability $p_j$ for each $f_j \in F = \{f_j\}_{j=1,...,k}$ according to Eq. (1).

Additionally, creating manually crafted templates takes time and experience and is possibly sub-optimal, failing to retrieve facts that the LM does know (Jiang et al., 2020). Inspired by relation extraction methods (Hearst, 1992), hand-built Hearst patterns such as “Y including X (Cities including Madrid or Barcelona)”, we create eight more lexico-syntactic templates to improve and stabilize the classification performance[3].

### 3.2 Distant Supervision Refinement

With a discipline-aware dictionary, we can generate distantly supervised labels by matching with the

[3] See more templates in Appendix A.3

**Algorithm 1** Dic-Matching with Academic Discipline

```
Input: Course Corpus $D = \{C_i\}_{i=1}^{n}$, where $C_i = \{S_i, F_i\}$; Dictionary $T = \{c_i\}_{i=1}^{m}$; $K$ number of Top-K;
  for each course $C_i = \{S_i, F_i\}$ in $D$ do
    for each video subtitles $v_{ij}$ in $S_i = \{v_{ij}\}$ do
      $X_m = [x_1, x_2, ...x_N] \leftarrow$ tokenize $v_{ij}$
      BIO tag potential concept using POS,RE:
      $D_m^{pot} = [d_1, d_2, ..., d_N]$
      for potential concept $pc_i$ tagged in $D_m^{pot}$ do
        if $pc_i \in T$ and
        (top-$K$ fields of $pc_i$) $\cap F_i \neq \emptyset$ then
          Tag BI to $pc_i$ tokens
        else
          Tag O to $pc_i$ tokens
        end if
      end for
      Academic discipline related:
      $D_m = [d_1, d_2, ..., d_N]$
    end for
  end for
Output: Distantly supervised labels $\{(X_m, D_m)\}_{m=1}^{M}$
```

top-$K$-related disciplines[4] in the ranked list from Eq. (1). This way, we can have a much higher precision by explicitly removing unrelated noisy annotations.

The entire Dic-Matching with academic discipline process is described in Algorithm 1. The input subtitles are first tokenized and annotated with part-of-speech (POS) tags. Next, we employ the regular expression (RE) by only keeping nouns to handle the noise challenge and mining more noun phrases to address the incomplete challenge, as illustrated in Appendix A.2. Finally filtering out unrelated disciplines, we have $\{(X_m, D_m)\}_{m=1}^{M}$ as distantly supervised data, where $X_m = [x_1, x_2, \ldots, x_N]$, composed of $N$ tokens, $D_m = [d_1, d_2, \ldots, d_N]$, based on the BIO schema (Li et al., 2012). Specifically, the first token of a concept mention is labeled as B; other tokens inside that concept mention are labeled as I; the non-concept tokens are labeled as O.

### 3.3 Discipline Embedding Self-training

We adapt the PLMs to the sequence labeling tasks with the distant labels and self-training approach to iteratively deal with the noisy annotations meantime training a new integrated embedding based on the concept discipline distribution to implicitly enhance model discipline-aware capability. Then we can employ other advanced DS-NER approaches, such as co-training and PUL.

#### 3.3.1 Discipline Embedding Model

At the pre-process of the dictionary, for each concept $c_i$, we calculate its distribution in all academic disciplines according to Eq. (1), denoted as $U_{c_i} = [p_1, p_2, \ldots, p_{|F|}]$. To introduce the discipline feature, each token $x_j$ of the input $X_m = [x_1, x_2, \ldots, x_N]$ is encoded as $E_j$ by adding its discipline distribution to BERT word embedding if $x_j$ is labeled as belonging to one of concept $c_i$ in the dictionary:

$$E_j = \begin{cases} Encoder(x_j) + U_{c_i}W & x_j \in c_i \\ Encoder(x_j) & x_j \notin any\, c_i \end{cases} \tag{2}$$

Where $d_h$ is a hidden dimension of the encoder, and $W \in R^{|F| \cdot d_h}$ is trainable parameters. We use BERT (Devlin et al., 2018) as our Encoder to learn the sequence representation. This way, external academic field features are integrated into the embedding, enhancing model discipline-aware capability (Figure 3).

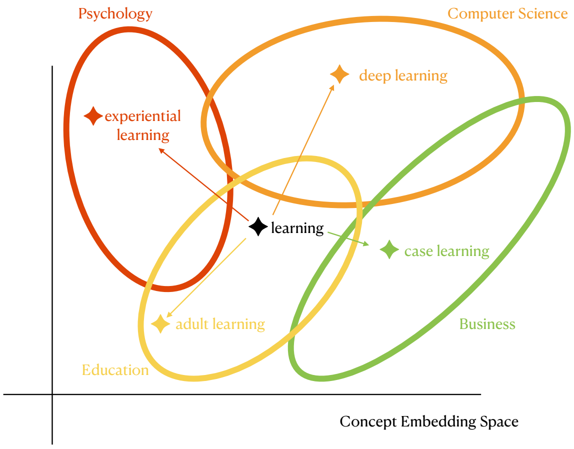


Figure 3: An example of external academic discipline feature promoting model discipline-aware capability.

Straightforward, we use $f(\cdot; \theta)$ to denote our model parameterized by $\theta$, which is a token-wise classifier on top of a pre-trained BERT. $f_{n,c}(\cdot; \cdot)$ denotes the probability of the $n$-th token in $X_m$ belonging to the $c$-th class from the BIO schema. The model will be learned by minimizing the cross

[4] $K$ is experimentally set to 2.

entropy loss $\mathcal{L}(\theta)$ over $\{(X_m, D_m)\}_{m=1}^{M}$:

$$\mathcal{L}(\theta) = \frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N} -\log f_{n,d_{m,n}}(X_m;\theta) \tag{3}$$

### 3.3.2 Teacher-Student Self-training

Following Liang et al. (2020); Meng et al. (2021); Zhang et al. (2021b); Liu et al. (2021a), we employ the teacher-student self-training strategy because it selects high-confidence and consistent predictions as pseudo labels from the teacher model and then uses them to guide the training of the student model, which removes the noisy labels iteratively. We adopt two advanced self-training DS-NER approaches. One is based on Zhang et al. (2021b), aimed at high-precision performance, which jointly trains two teacher-student networks and confirms its effectiveness and robustness in dealing with the label noise.

The other is inspired by Peng et al. (2019), aimed at high-recall performance, which introduces PUL as it can unbiasedly and consistently estimate the task loss. We apply the binary label assignment mechanism for using this algorithm by mapping "O" to 0 and "B", "I" to 1. Finally, we get positive set $D_m^+ = [d_{m,1}, ..., d_{m,|D^+|}]$ and unlabeled set $D_m^u = [d_{m,1}, ..., d_{m,|D^u|}]$ from the original distantly supervised labels $D_m = [d_{m,1}, d_{m,2}, ..., d_{m,N}]$. The PUL training loss is defined by:

$$\widehat{\mathcal{L}}(\theta) = \gamma \cdot \pi_p \widehat{\mathcal{L}}_p^+(\theta) + max\{0, \widehat{\mathcal{L}}_u^-(\theta) - \pi_p \widehat{\mathcal{L}}_p^-(\theta)\} \tag{4}$$

where

$$\widehat{\mathcal{L}}_p^+(\theta) = \frac{1}{M}\frac{1}{|D^+|}\sum_{m=1}^{M}\sum_{d=1}^{|D^+|} -\log f_{d,1}(X_m;\theta)$$

$$\widehat{\mathcal{L}}_p^-(\theta) = 1 - \widehat{\mathcal{L}}_p^+(\theta)$$

$$\widehat{\mathcal{L}}_u^-(\theta) = \frac{1}{M}\frac{1}{|D^u|}\sum_{m=1}^{M}\sum_{d=1}^{|D^u|} -\log f_{d,0}(X_m;\theta)$$

and $\pi_p$ is the ratio of positive concept words within $D_u$. A class weight $\gamma$ is introduced to deal with the class imbalance problem ($\pi_p$ is very small). As a whole, in this training strategy, the parameters of the student model $\theta^*$ are learned by the combination of the cross entropy loss (Eq. (3)) and the PUL loss (Eq. (4)):

$$\theta^* = \underset{\theta}{argmin}(\mathcal{L}(\theta) + \beta \cdot \widehat{\mathcal{L}}(\theta)) \tag{5}$$

where a parameter $\beta$ is introduced to balance these two loss functions.

# 4 Experiments

## 4.1 Experimental Settings

### 4.1.1 Dataset

We provide a new dataset spanning 20 academic disciplines, which can be used to benchmark distantly supervised methods for course concept extraction task in MOOCs. Based on MOOCCube (Yu et al., 2020a), the input includes two parts: (1) an expert-checked dictionary with over 100k course concepts from CNCTST[5], and (2) a subtitle corpus from 315 courses with $167,496$ unlabeled character sequences on average per course. The test set contains 522 expert-annotated sentences from 17 courses with $15,375$ discipline-related course concepts. All data is from XuetangX[6], one of the largest MOOC websites in China, so the dataset is in the Chinese language. More details of the dataset can be found in Appendix A.5.

### 4.1.2 Baselines and Evaluation Metrics

We compare our method with several competitive baselines from three aspects and use Precision (P), Recall (R), and F1 score as the evaluation metrics.

**Dic-Matching Methods**. We construct different Dic-Matching (DM) methods for comparison, including (i) **DM**: it is a simple string matching with a greedy search algorithm to find the longest matching strings in sentences; (ii) **DM(AD-LM)** : it adopts the matching strategy proposed in Algorithm 1; (iii) **DM(AD-human)**: it is a variation of AD-LM that replaces the discipline classification results from GLM with ones from CNCTST expert annotations.

**Fully-supervised Method**. We also construct fully-supervised methods for comparison. **FLAT** (Li et al., 2020): For Chinese NER, it converts the lattice structure into a flat structure consisting of spans to handle word segmentation in the Chinese language.

**Distantly-supervised Methods**. The state-of-the-art self-training DS-NER methods are as follows. (i) **SCDL** (Zhang et al., 2021b): It explores more helpful information from the mislabeled data by a devised co-training paradigm based on self-training. (ii) **RoSTER** (Meng et al., 2021): A self-training method that uses contextualized augmentations created by pre-trained language models to improve the model's generalization ability. (iii) **BOND**(Liang

[5] http://www.cnterm.cn/

[6] https://next.xuetangx.com/

et al., 2020): A two-stage framework that trains a RoBERTa model on distantly-labeled data with early stopping in the first stage and improves the model fitting with a teacher-student framework to iteratively self-train the model in the second stage.

### 4.1.3 Implementation Details

For concept classification task, we apply the General Language Model (Du et al., 2022), which is capable of handling variable-length blank. We use the pre-trained BERT-wwm-ext model (Cui et al., 2020) as the backbone for our method and other distantly-supervised baselines. The maximum sequence length of our dataset is set to be 512 tokens. The max training epoch is 30, and the batch size is 4. We use Adam (Kingma and Ba, 2014) as the optimizer, and the learning rate is $10^{-5}$. The confidence threshold $\gamma$ is 0.9 for the co-training strategy while 0.7 for the PUL strategy with the purpose of high-recall performance. More implementation details can be found in Appendix A.4.

## 4.2 Experimental Results

**Overall Results.** Table 2 shows the overall results of different methods on our MOOCs test set. Our DS-MOCE framework with two self-training strategies achieves the best performance among distantly-supervised methods. Specifically, (1) the proposed Dic-Matching method with academic discipline refines the distant labels by improving precision significantly; (2) **DS-MOCE(co)** reports a high precision with 7% absolute F1 score improvement over the best performing baseline model BOND, demonstrating the superiority of our proposed Dic-Matching with academic discipline method and self-training approach; (3) **DS-MOCE(PUL)** consistently outperforms other distantly-supervised methods with a higher recall and reasonable precision, showing more robustness to the issue of incomplete labeling.

As we have discussed, the Dic-Matching method suffers from extremely low precision and low recall in MOOCs for its diversity, which dramatically hurts the performance of the distantly supervised baselines and limits the model fitting ability in fully supervision.

**Discipline Classification Results.** Through the comparison of **DM(AD-human)** and **DM(AD-LM)** in Table 2, we find that the academic discipline classification result from GLM outperforms that from expert annotations during Dic-Matching, showing the robustness of our designed classification step for transferring PLMs knowledge to the dictionary. On the contrary, human annotations suffer from missing, incorrect, and out-of-date classifications.

| Method | P | R | F1 |
|---|---|---|---|
| **Dic-Matching** | | | |
| DM | 12.50 | 25.84 | 16.85 |
| DM(AD-human) | 22.95 | 17.38 | 19.78 |
| DM(AD-LM) | 34.59 | 15.40 | 21.31 |
| **Distant-Sup.** | | | |
| SCDL | 34.59 | 21.16 | 26.26 |
| RoSTER | 35.40 | 26.70 | 30.40 |
| BOND | 32.37 | 44.78 | 37.58 |
| **Our DS-MOCE** | | | |
| DS-MOCE(co) | **81.93** | 30.82 | **44.79** |
| DS-MOCE(PUL) | 34.53 | **49.34** | 40.62 |
| **Sup.** | | | |
| FLAT | 56.08 | 57.17 | 56.62 |

Table 2: Overall results (%) on our MOOCs test set.

Moreover, we evaluate the pre-process concept classification task using the Mean Average Precision (MAP), a metric in information retrieval for evaluating ranked lists. Table 3 shows some example results using different templates. (See more templates in Appendix A.3). The first example is based on experience and the rest are Hearst patterns, showing better and more stable performance. We finally use the best-performing template in the following parts.

| Template | MAP |
|---|---|
| $[concept]$ belongs to $[MASK]$ | 51.35 |
| $[concept]$, a concept of $[MASK]$ | 58.44 |
| $[MASK]$, especially $[concept]$ | 58.89 |
| $[MASK]$ including $[concept]$ | 59.95 |

Table 3: Results (%) of different templates.

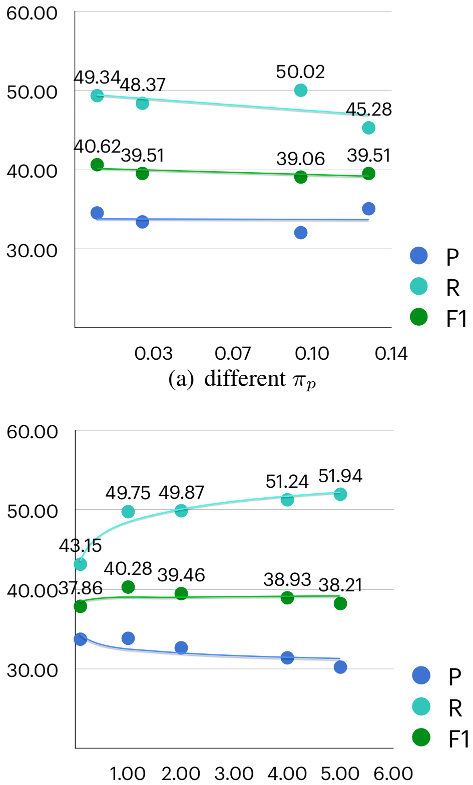
**Ablation Study.** To evaluate the influence of each component, we conduct the following ablation study for further exploration by removing one component at a time: (1) do not adopt Alg. 1 and use the Dic-Matching method when generating distantly supervised labels. (2) only use BERT Encoder without adding academic discipline embedding; (3) do not perform self-training; (4) do not perform co-training for **DS-MOCE(co)** and only use cross-entropy loss in Eq. (3) without adding PUL loss in Eq. (4) for **DS-MOCE(PUL)**. The results are shown in Table 4. It can be seen that w/o Alg.

| Method | P | R | F1 |
|---|---|---|---|
| **DS-MOCE(co)** | 81.93 | 30.82 | 44.79 |
| w/o Alg. 1 | 14.15 | 34.28 | 20.04 |
| w/o embedding | 34.59 | 21.16 | 26.26 |
| w/o self-train | 60.07 | 27.38 | 37.61 |
| w/o co | 67.63 | 30.74 | 42.47 |
| **DS-MOCE(PUL)** | 34.53 | 49.34 | 40.62 |
| w/o Alg. 1 | 17.52 | 50.26 | 25.98 |
| w/o embedding | 34.30 | 48.79 | 40.28 |
| w/o self-train | 14.33 | 40.66 | 21.19 |
| w/o PUL | 32.02 | 36.63 | 34.17 |

Table 4: Ablation study results (%).

| Discipline | ratio | P | R | F1 |
|---|---|---|---|---|
| Philosophy | 0.05 | 80.65 | 11.57 | 20.24 |
| CS | 0.27 | 84.16 | 47.87 | 61.03 |
| Mathematics | 0.16 | 92.89 | 48.50 | 63.73 |
| Medicine | 0.16 | 89.38 | 22.90 | 36.46 |

Table 5: DS-MOCE(co) results (%) on courses from different disciplines. The ratio is calculated by ( # of concept words) / (# of words of subtitles). See specific course names in Appendix A.5.1.



(a) different $\pi_p$

(b) different $\beta$

Figure 4: Parameter study results (%) of DS-MOCE(PUL).

1 refinement and w/o embedding for both strategies lead to worse performance than the full model, confirming the necessity of considering discipline features in MOOCs. Removing the self-training or co-training component also reduces performance, showing its importance in DS-MOCE(co) of denoising learning because false-negative labels can be explored via peer model or another network iteratively. Without PUL, the recall value decreases sharply, which validates the effectiveness of introducing PUL to tackle the incomplete challenge.

**Parameter Study.** Before discussing the parameter study of $\pi_p$ defined in Eq. (4), we first calculate the true value of $\pi_p$ = (# of concept words) / (# of words of the training set) in our dataset, with a 0.1002 result. Then we train the proposed model DS-MOCE(PUL) with different estimated $\pi_p$, and evaluate its performance on the test set. From Figure 4(a), we can see that although the highest recall is achieved by setting $\pi_p$ = 0.1, most closely to the true value, the variation of results across different $\pi_p$ is relatively tiny. This motivates us to use a proper estimated value of $\pi_p$ to deal with the diversity of MOOCs where courses from different disciplines have incongruous and unknown $\pi_p$ values. Therefore, we set $\pi_p = 0.01$ for DS-MOCE(PUL) to achieve a high recall and a higher F1 score. Besides, we set $\beta = 1$ in Eq. (5) throughout our experiments without further illustration, according to Figure 4(b).

**Different Discipline Analysis.** We analyze that the diversity of MOOCs academic disciplines accounts for more noisy and incomplete annotations in distantly supervised MOOCs. As a result, we select some courses from different disciplines and use DS-MOCE(co) framework to perform prediction on these courses. From Table 5, we discover that (1) the intensive and appropriate terminological concepts in formal and applied science, such as CS and Mathematics, bootstrap the model with its high-recall predictions that benefit the model's generalization; (2) the sparse distribution (low concept ratio) in Humanities and Social Science Philosophy makes it uncertain about selecting tokens to train a robust model; (3) excessive terminological concepts (nested structure and long formulas) in some professions, such as Chemistry and Medicine, amplify the issue of incomplete annotation, where several concept extraction methods have been developed specifically to handle this problem (Wang et al., 2021; Fu et al., 2020).

| | |
|---|---|
| **Sentence # 1** | 操作系统的功能是在用户态和硬件之间，......<br>The function of the operating system is ... between the user state and the hardware. |
| **DM** | 操作系统的功能是在用户态和硬件之间，...... |
| **DM(LM)** | 操作系统的功能是在用户态和硬件之间，...... |
| **SCDL** | 操作系统的功能是在用户态和硬件之间，...... |
| **RoSTER** | 操作系统的功能是在用户态和硬件之间，...... |
| **BOND** | 操作系统的功能是在用户态和硬件之间，...... |
| **DS-MOCE(co)** | 操作系统的功能是在用户态和硬件之间，...... |
| **Sentence # 2** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。<br>Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |
| **DM** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。 |
| **DM(LM)** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。 |
| **DS-MOCE(co)** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。 |
| **DS-MOCE(PUL)** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。 |

Table 6: Case studies between DS-MOCE(co) and baselines of the first sentence; DS-MOCE(co) and DS-MOCE(PUL) of the second sentence. Golden labels are marked in orange. Noisy labels are marked in red and incomplete in blue. See the corresponding English illustration in Appendix A.6.

**Case Study.** Finally, we perform a case study to understand the advantage of DS-MOCE(co) with a concrete example in Table 6.

Besides, we select another case study to demonstrate why DC-MOCE(PUL) is provided in our work. The extremely high precision accounts for the F1 score increment of the DS-MOCE(co) framework, but low recall leads to more missing tokens. Consequently, aimed at improving recall, we design the DC-MOCE(PUL) as an alternative option by sacrificing the precision properly. Finally, DS-MOCE(co) with high-precision and DS-MOCE(PUL) with high-recall can be applied in different real-world scenarios.

To help our model's behaviors be understood and applied to real-world applications, we suggest: (1) For DS-MOCE(co) with high-precision performance, it is better to apply it to the downstream tasks that acquire accurate concepts but ignore the coverage, such as course concept recommendation and AI-driven robot assistant; (2) For DS-MOCE(PUL) with high-recall performance, it is better to apply it to scenarios where there is surplus human labor available for corrections, and where there is a need to recall as many course concepts as possible, such as in MOOC knowledge graph construction.

## 5 Related Work

**Distantly Supervised NER**. Our work is more closely related to distantly supervised NER, where the primary research focuses on coping with the noise and incomplete annotations problem.

Several new training paradigms have been proposed along the denoising line, such as Reinforcement learning (Yang et al., 2018), AutoNER (Shang et al., 2018) with a new tagging scheme "tie or break", Hypergeometric Learning (Zhang et al., 2021a) and Bagging-based active learning with negative sampling (Lee et al., 2016). Along the incomplete mining line, a direct solution is concept expansion (Yu et al., 2019, 2020b; Wang et al., 2019), which finds new candidates and ranks them to expand the set based on the seed set with figurative elements. AdaPU (Peng et al., 2019) and Conf-MPU (Zhou et al., 2022) are developed to address the incomplete challenge by formulating the task as a positive-unlabeled learning problem. Besides, many studies (Yang et al., 2018; Shang et al., 2018) attempt to modify the standard CRF to partial annotation CRF to consider all possible labels for unlabeled tokens. However, these works do not work well in MOOCs where severe low-precision and low-recall problems have been reported previously.

**Course Concept Extraction**. Our study is also relevant to course concept extraction, which is related to keyphrase extraction (Hasan and Ng, 2014) in the information retrieval domain. The well-known methods such as tf-idf (Ramos et al., 2003), co-occurrence (Mihalcea and Tarau, 2004), and PositionRank (Florescu and Caragea, 2017) are frequently used in unsupervised automatic keyphrase extraction. However, the low-frequency (i.e., appearing only once or twice in the subtitles) feature of keyphrases in MOOCs makes statistical information less useful (Pan et al., 2017). Therefore, Pan et al. (2017) develop a graph-based propagation algorithm, and Albahr et al. (2021) design a novel unsupervised cluster-based approach to address the low-frequency problem in keyphrases extraction from MOOCs.

DS-MOCE also benefits from distributed representations of words, namely word embeddings (Mikolov et al., 2013) to learn academic discipline representations for concepts from the dictionary, which has been employed in Wang et al. (2018); Wu et al. (2022).

## 6 Conclusion and Future Work

In this paper, we attribute the increased noise and incomplete challenges of distantly supervised course concept extraction in MOOCs to the limited dictionary and diverse MOOCs. To tackle these challenges, we propose a three-stage framework DS-MOCE, which handles the unrelated noise through Dic-Matching refinement and discipline-embedding model training, and leverages the power of pre-trained language models for dictionary empowerment and incomplete mentions mining. We also provide an expert-labeled dataset spanning 20 academic disciplines. Experimental results show that DS-MOCE is highly effective, outperforming the state-of-the-art distantly supervised methods. Although achieving significant improvement, course concept extraction in MOOCs is still nontrivial. In the future, we plan to design a more robust training method to jointly deal with severe noisy and incomplete issues and apply it to other real-world open domains.

## 7 Ethic Consideration

We provide an expert-labeled dataset spanning 20 academic disciplines, which contains 522 expert-annotated sentences from 17 courses with $15,375$ course concepts. We define the 20 academic disciplines according to *Discipline Doctor and Master Degree and postgraduate training, the professional directory* issued by the Ministry of Education of the People's Republic of China[7]. The course corpus is collected from an open-source database MOOCCube (Yu et al., 2020a)[8]. The dictionary is collected from CNCTST[9] with expert-checked 100k course concepts. The annotated sentences in the test set are from an expert from the Education Department in our university, which may have limitations but missing criteria in MOOCs means that we can accept this human bias. The annotator is a voluntary participant who was aware of any risks of harm associated with their participation and had given their informed consent. To lighten the burden of the annotator, we first use unsupervised methods, such as tf-idf, to give a rough annotation result for each course, randomly selected from XuetangX[10]. Then the annotator marks mentions of high-quality course concepts based on that. More details of the dataset can be found in Appendix A.5.

## 8 Limitations

Although we conducted extensive experiments, the exploration scope of this work has some limitations: (1) All data is from one of the largest MOOC websites in China, so the dataset is in the Chinese language, which limits the linguistic features covered in our analyses. We will add comprehensive corpora from other MOOC platforms with various languages such as English, Japanese, French, and so on to enhance the availability and coverage of our dataset. (2) We present two models with high-precision and high-recall behaviors. The severe noisy and incomplete issues could not be coped with simply by combining two technical methods (i.e. co-training and PUL). A more robust training method should be proposed to jointly achieve better overall performance. We encourage future works to address these limitations and get more comprehensive analysis results.

## Acknowledgements

This work is supported by a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003). This work is also supported by the NSFC Youth Project (62006136).

## References

Abdulaziz Albahr, Dunren Che, and Marwan Albahar. 2021. A novel cluster-based approach for keyphrase extraction from mooc video lectures. *Knowledge and Information Systems*, 63(7):1663–1686.

[7] http://www.moe.gov.cn/srcsite/A22/moe_833/200512/t20051223_88437.html

[8] http://moocdata.cn/data/MOOCCube

[9] http://www.cnterm.cn/

[10] https://next.xuetangx.com

Graham Butt and Ann Lance. 2005. Secondary teacher workload and job satisfaction: do successful strategies for change exist? *Educational Management Administration & Leadership*, 33(4):401–422.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Corina Florescu and Cornelia Caragea. 2017. PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, Vancouver, Canada. Association for Computational Linguistics.

Sunyang Fu, David Chen, Huan He, Sijia Liu, Sungrim Moon, Kevin J Peterson, Feichen Shen, Liwei Wang, Yanshan Wang, Andrew Wen, et al. 2020. Clinical concept extraction: a methodology review. *Journal of biomedical informatics*, 109:103526.

Kazi Saidul Hasan and Vincent Ng. 2014. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, Baltimore, Maryland. Association for Computational Linguistics.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sunghee Lee, Yeongkil Song, Maengsik Choi, and Harksoo Kim. 2016. Bagging-based active learning model for named entity recognition with distant supervision. In *2016 International conference on big data and smart computing (BigComp)*, pages 321–324. IEEE.

Qi Li, Haibo Li, Heng Ji, Wen Wang, Jing Zheng, and Fei Huang. 2012. Joint bilingual name tagging for parallel corpora. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1727–1731.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online. Association for Computational Linguistics.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. Bond: Bert-assisted open-domain named entity recognition with distant supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064.

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021a. Noisy-labeled ner with confidence estimation. *arXiv preprint arXiv:2104.04318*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Weiming Lu, Yangfan Zhou, Jiale Yu, and Chenhao Jia. 2019. Concept extraction and prerequisite relation learning from educational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9678–9685.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Huzaifah Mohd Salamon, Nazmona Mat Ali, Suraya Miskon, and Norasnita Ahmad. 2016. Initial recommendations of moocs characteristics for academic discipline clusters. 87:204–213.

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.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. *CoRR*, abs/1906.01378.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*.

Zhengyang Song, Jie Tang, Tracy Xiao Liu, Wenjiang Zheng, Lili Wu, Wenzheng Feng, and Jing Zhang. 2021. Xiaomu: an ai-driven assistant for moocs. *Science China Information Sciences*, 64(6):1–3.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Xiaochen Wang, Wenzheng Feng, Jie Tang, and Qingyang Zhong. 2018. Course concept extraction in mooc via explicit/implicit representation. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 339–345.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly supervised biomedical named entity recognition with dictionary expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 496–503.

Zhijie Wu, Jia Zhu, Shi Xu, Zhiwen Yan, and Wanying Liang. 2022. Ltwnn: A novel approach using sentence embeddings for extracting diverse concepts in moocs. In *Australasian Joint Conference on Artificial Intelligence*, pages 763–774. Springer.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169.

Jifan Yu, Gan Luo, Tong Xiao, Qingyang Zhong, Yuquan Wang, Wenzheng Feng, Junyi Luo, Chenyu Wang, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jie Tang. 2020a. MOOCCube: A large-scale data repository for NLP applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3135–3142, Online. Association for Computational Linguistics.

Jifan Yu, Chenyu Wang, Gan Luo, Lei Hou, Juanzi Li, Jie Tang, Minlie Huang, and Zhiyuan Liu. 2020b. Expanrl: Hierarchical reinforcement learning for course concept expansion in moocs. In *Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 770–780.

Jifan Yu, Chenyu Wang, Gan Luo, Lei Hou, Juanzi Li, Jie Tang, and Zhiyuan Liu. 2019. Course concept expansion in moocs with external knowledge and interactive game. *arXiv preprint arXiv:1909.07739*.

Wenkai Zhang, Hongyu Lin, Xianpei Han, Le Sun, Huidan Liu, Zhicheng Wei, and Nicholas Yuan. 2021a. Denoising distantly supervised named entity recognition via a hypergeometric probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14481–14488.

Xinghua Zhang, Bowen Yu, Tingwen Liu, Zhenyu Zhang, Jiawei Sheng, Mengge Xue, and Hongbo Xu. 2021b. Improving distantly-supervised named entity recognition with self-collaborative denoising learning.

Kang Zhou, Yuepei Li, and Qi Li. 2022. Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7198–7211, Dublin, Ireland. Association for Computational Linguistics.

# A Appendix

## A.1 Case Explanations of Limited Dictionary & Diverse MOOCs

**The limited dictionary.** It is expensive and time-consuming to expand or tailor a dictionary to every specific domain because of MOOCs rapid growth and criteria missing, ending up with out-of-dictionary and low-quality concept problems. We categorize two types of low-quality course concepts in the dictionary. The first type is not specific enough, missing prefixes and suffixes. The second type is unigram concepts with many extended meanings, which end up with false-positive labels.

**The diverse MOOCs.** Compared with other benchmark datasets, Table 1 illustrates that the number of concept types is inversely proportional to the distantly matching performance. As shown in Table 1, where BC5CDR (Shang et al., 2018) is restricted to the biomedical domain, a domain-specific dictionary with a corpus-aware dictionary tailoring method can achieve higher precision and reasonable recall. MOOCs can span 20 or even more academic disciplines. During label generation, unrelated concept annotations would produce more false-positive noise.

Besides, the characteristics among varied disciplines are different. Most of the time, the concept distribution in humanities and social science is sparse, while in formal science is dense. According to the statistics, the concept proportion of contents in one psychology course is 0.0163, whereas in one computer science course is 0.1. The uneven concept distribution may lead to a matching bias toward the concept-intensive academic discipline. Furthermore, in Chinese, homonyms are more likely to appear in humanities and social science, where words share the same characters and pronunciations but have different meanings. For example, in Philosophy, the debate of right and wrong makes "right" annotations correct. However, "right", as a high-frequency phrase, is easily annotated in other contexts, producing false-positive labels. The ambiguity of homonyms makes it difficult to extract the correct meaning concept in these domains.

## A.2 Regular expression in Distant Supervision Refinement

During distant supervision refinement, we employ the following regular expression, introduced by Luo et al.[11], only keeping nouns and noun phrases to remove the apparent incorrect POS noise and mining more incomplete annotations by connecting two nouns with @.

$$(@(([av]?n[rstz]?)|l|a|v))*(@(([av]?n[rstz]?)|l))$$

| Template | MAP |
|---|---|
| $[concept]$, a concept of $[MASK]$ | 58.44 |
| $[MASK]$ such as $[concept]$ | 56.77 |
| $[MASK]$ including $[concept]$ | 59.95 |
| $[concept]$ and other $[MASK]$ | 54.63 |
| $[concept]$ or other $[MASK]$ | 54.32 |
| $[concept]$ which is known as $[MASK]$ | 54.57 |
| $[MASK]$, especially $[concept]$ | 58.89 |
| like $[MASK]$, $[concept]$ | 54.87 |

Table 7: Results (%) of Hearst Pattern Templates.

## A.3 Templates Results

All eight template results based on Hearst Pattern are shown in Table 7.

## A.4 Baselines Settings

For fully-supervised methods, we use 3/4 of the test set for model training and the rest for evaluation. Fairly, the following distantly supervised methods use the distantly-labeled training set obtained from Dic-Matching(**AD-GLM**).

- **SCDL.** We use the authors' released code: `https://github.com/AIRobotZhang/SCDL`. Because our test set is in the Chinese language, we change the basic model to the same pre-trained BERT-wwm-ext model with our method. We train the model for 30 epochs with a batch size of 8. The other hyperparameters are set to default values.
- **RoSTER.** We follow the officially released implementation from the authors: `https://github.com/yumeng5/RoSTER`. Similarly, we modify the backbone model from RoBERTa-base to the same one with our method. The epoch number is set to 3, 3, and 7 for noise-robust training, ensemble model training, and self-training, respectively. We train five models with 2000 intervals of noise-robust training and 1000 of self-training with

[11] a parent with number CN201911140653.9

| Metrics | results |
|---|---|
| number of course | 17 |
| Avg. number of video | 12.06 |
| Avg. length of subtitles | 15740.71 |
| Avg. number of related disciplines | 1.82 |
| Avg. number of concepts | 904.41 |
| Max. number of concepts | 5174 |
| Avg. length of concept | 2.39 |

Table 8: Test set information.

a batch size of 8. The rest hyperparameters are the same as the default values.

- **BOND.** We use the authors' released code: https://github.com/cliang1453/BOND/. Also, we choose the pre-trained BERT-wwm-ext model as the backbone model. The early stopping step of the student model is set to 100k. The other hyperparameters are set to default values.

## A.5 Dataset Statistic

### A.5.1 Test Set Annotation

We select 17 courses from the course corpus spanning these disciplines and ask an expert to annotate each sentence as our test sets. The more detailed statistics are shown in Table 8. During analysis of different discipline, we choose *Introduction to the Classical Works of Chinese Philosophy* for Philosophy; *Machine Learning for Big Data* for Computer Sciences(CS); *Finite Element Analysis and Applications* for Mathematics; *Pathology* for Medicine.

### A.5.2 Dictionary Information

We created our dictionary with 20 academic disciplines by developing the resource from MOOCCube(Yu et al., 2020a) based on its concept taxonomy from CNCTST. Then according to *Discipline Doctor and Master Degree and postgraduate training, the professional directory* issued by the Ministry of Education of the People's Republic of China[12], we show the prescribed 20 academic disciplines and the distribution of concepts that filtered and mapped from MOOCCube in Table 9.

## A.6 Case studies in English

To make the case study more vivid, we highlight the corresponding English word in different colors in Table 10. Considering contextual differences between the two languages, there are some missing tokens in English.

[12] http://www.moe.gov.cn/srcsite/A22/moe_833/200512/t20051223_88437.html

| Academic Discipline | Abbreviation | in Chinese | #concepts |
|---|---|---|---|
| Philosophy | Phi. | 心理学 | 2136 |
| Education | Edu. | 教育学 | 2947 |
| Linguistics and languages | Lin. | 语言学 | 2909 |
| History | His. | 世界历史 | 4021 |
| Mathematics | Mat. | 数学 | 7876 |
| Physics | Phy. | 物理学 | 4273 |
| Chemistry | Che. | 化学 | 6909 |
| Mechanics | Mec. | 力学 | 1119 |
| Mechanical Engineering | ME | 机械工程 | 18011 |
| Materials Science | MS | 材料科学技术 | 6923 |
| Electrical Engineering | EE | 电气工程 | 5000 |
| Computer Science | CS | 计算机科学技术 | 4906 |
| Architecture | Arc. | 建筑学 | 5305 |
| Marine Engineering | ME | 船舶工程 | 2333 |
| Aeronautical | Aer. | 航天科学技术 | 4213 |
| Aviation | Avi. | 航空科学技术 | 2236 |
| Agriculture | Agr. | 农学 | 2248 |
| Medicine | Med. | 医学 | 10346 |
| Business | Bus. | 管理科学技术 | 7473 |
| Immunology | Imm. | 免疫学 | 1564 |

Table 9: 20 Academic disciplines and the concept distribution.

| | |
|---|---|
| **Sentence # 1** | 操作系统的功能是在用户态和硬件之间，......<br>The function of the operating system is ... between the user state and the hardware. |
| **DM** | The function of the operating system is ... between the user state and the hardware. |
| **DM(LM)** | The function of the operating system is ... between the user state and the hardware. |
| **SCDL** | The function of the operating system is ... between the user state and the hardware. |
| **RoSTER** | The function of the operating system is ... between the user state and the hardware. |
| **BOND** | The function of the operating system is ... between the user state and the hardware. |
| **DS-MOCE(co)** | The function of the operating system is ... between the user state and the hardware. |
| **Sentence # 2** | 传染性疾病是由病毒，细菌，原生动物和寄生虫等等一系列的微生物产生。<br>Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |
| **DM** | Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |
| **DM(LM)** | Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |
| **DS-MOCE(co)** | Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |
| **DS-MOCE(PUL)** | Infectious diseases are produced by a range of microorganisms such as viruses, bacteria, protozoa and parasites. |

Table 10: Table 6 illustration in English. Case studies between DS-MOCE(co) and baselines of the first sentence; DS-MOCE(co) and DS-MOCE(PUL) of the second sentence. Golden labels are marked in orange. Noisy labels are marked in red and incomplete in blue.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

□ A1. Did you describe the limitations of your work?
*Left blank.*

□ A2. Did you discuss any potential risks of your work?
*Left blank.*

□ A3. Do the abstract and introduction summarize the paper's main claims?
*Left blank.*

□ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B □ Did you use or create scientific artifacts?

*Left blank.*

□ B1. Did you cite the creators of artifacts you used?
*Left blank.*

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Left blank.*

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Left blank.*

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Left blank.*

### C □ Did you run computational experiments?

*Left blank.*

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Left blank.*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Left blank.*

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Left blank.*

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Left blank.*

**D □ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Left blank.*

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Left blank.*

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Left blank.*

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Left blank.*

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Left blank.*