NLPCC2024 Shared Task4

Chinese Essay Discourse Logic Evaluation and Integration

Guideline

Yuhao Zhou(1), Hongyi Wu(1), Xinshu Shen(1), Man Lan(1, 3), Yuanbin Wu(1),
Xiaopeng Bai(2), Shaoguang Mao(4), Tao Ge(4), Yan Xia(4)

（1.College of Computer Science and Technology, East China Normal University, Shanghai 200062;

2.Department of Chinese Language and Literature, East China Normal University, Shanghai 200241;

3.Institute of AI EDUCATION, East China Normal University, Shanghai 200062;

4.Microsoft Research Asia, Beijing 100080）

Contact: Yuhao Zhou (yhzhou@stu.ecnu.edu.cn)

April 13th, 2024

Table of Contents

1. Background

In accordance with the essay grading criteria of the Shanghai middle school examination, the coherence of thought and structure is a critical component in assessing the level of essay writing. Coherence assessment of essays necessitates not only the analysis of semantic flow within sentences but also an understanding of the logical relationships between sentences and the overarching logical semantic structure of the text. This understanding is vital for grasping the logic in long texts and generating complex documents. However, due to a shortage of large-scale, high-quality, comprehensive datasets for the evaluation of discourse logic, current studies have not systematically delved into the relationship between discourse logic, expressive capacity, and text quality. This limitation has also impeded the progress of AI-based essay correction technologies.

To address this challenge, the Institute of AI EDUCATION in East China Normal University, in collaboration with Microsoft Research Asia, the CubeNLP Laboratory of the Computer Science School, and the Department of Chinese Language and Literature at East China Normal University, have developed a Chinese Essay Discourse Coherence Corpus (**CEDCC**). In 2023, the CEDCC was used in the first "Middle and High School Essay Discourse Coherence Evaluation" task at the NLPCC, which garnered substantial interest. With 23 teams participating from prestigious universities, research institutions, and corporations, there were 82 submissions, receiving widespread acclaim.

This year, leveraging the NLPCC 2024 platform, we are organizing the second edition of the Essay Discourse Logic Evaluation and Integration Task. Compared to last year, the current evaluation task includes both an expanded investigation into the various dimensions affecting discourse coherence and an exploration of the interactions between these factors. Specifically, we have added segments for discourse logical error detection(**DLED**), topic coherence modeling(**TCM**), and discourse coherence feedback generation(**DCFG**). Our aim is to provide a high-quality resource for the evaluation of discourse coherence and a testing ground that will contribute to more profound exploration and advancement in the field of Chinese essay discourse coherence research.

2. Task Overview

2.1 Track 1. Discourse Logic Error Detection (DLED)

2.1.1 Task Description

Cohesion in discourse is an essential element in linguistic expression, involving the linkage between sentences and the smooth transitions between paragraphs. A high-quality essay should ensure clarity in logic and a distinct hierarchy between sentences and paragraphs. This task is dedicated to identifying and evaluating the fine-grained logical errors that impact discourse coherence. Through analysis based on the Chinese Essay Discourse Coherence Corpus (**CEDCC**), it examines the capabilities of existing technologies in detecting the coherence of essay discourse structures and explores the potential contributions of discourse logical structure in enhancing the accuracy of assessments.

2.1.2 Task Definition

Given an essay written by high school students, annotators are required to identify whether each sentence contains logical errors. This includes improper use of conjunctions, lack of logical relationships between contexts, unreasonable sentence breaks and deviating from the main topic.

The following are examples of fine-grained logical errors that affect the coherence of discourse.

错误类型	定义	文章示例	解释
偏题	文章有部分内容对中心思想有所偏离。	对人们来说，做事情，要做自己感兴趣的。什么是该做的，什么是不该做。在做事这一方面，每个人都是有人们自己的人生经历。\n 在人们的人生经历中，从小到大，做事是从我做起的。从出生至现在，人们小时候不懂事，非常喜欢乱做事，喜欢贪玩、捣蛋等，到了幼儿时期，比三岁更懂事，喜欢读书。到了六岁，读书的次数比之前越来越多，贪玩的次数越来越少。进入小学阶段，人们就开始学汉语拼音，开始学汉字。在人生的成长经历中，既有成功又有挫折。人们培养的好习惯越来越多，不好的越来越少。到了初中阶段，读书的次数越来越多。当初中成绩去考不好时，那就上不了高中，当高中成绩考不好时，就上不了大学。考完大学之后，会走向社会且找好工作，等等......\n 每个人做的事的结果是不一样的。人们做的事都要有兴趣是不一定正确的，对人们来说，人们每个能力有大有小，但是这点人都会有精神的，对于做事方面，有些事情是对人们感兴趣的。读书是对人们最感兴趣的，会让人们放松下来。\n 你从出生至现在，小时候是不肯读书且爱贪玩，长大之后，比之前懂事了，书就开始读的越来越多了。到了小学高年级阶段，书就读得越来越多了。到了初中阶段，书就读的越来越多，就越来越兴趣了。未来，到上高中、大学时期。书要反复读且要理解含义，就更会有兴趣了，下次要少玩电子产品中的游戏，用上网搜资料，这是不必要的。\n 做事是有时间的，大部分都不是一气呵成的。要有“少年强，则国强”的人生道路。	1. 文章主题用红色高亮部分标出；文章中偏题部分用黄色高亮部分标出； 2. 以本篇文章为例，文章主题为“人要做自己感兴趣的事情”，但文章有大部分和“兴趣”无关的内容，视作偏题。

		——《对人们做的事》	
连接词使用不当	在文章中连接词选用错误	太阳将光芒洒向大地，覆盖地上的一切，唯有一些不起眼的角落阴森悲凉。\n 那个角落也许就在人们心里，那是人心中不被光临了一丝黑暗，也那就在我们身边，一个光临了心灵黑暗的孩子在那里句泪水倾诉着。\n 漫步在阳光下的我们，也被阳光覆盖着，此时我们的心情随着阳光下的景物而愉悦。“更无法理解为什么会有泪水，但当我们孤独的一个人待在阳光之外时，我们何尝能回忆到那种愉悦，又流下了泪水。\n 阳光是什么？它可以是宇宙中天体放出了能量，也可以是一种性格特点，更可以是人们的生存寄托。\n 我们待在阳光下，不再漫步，我们会感到燥热、烦躁，景随情变，不再鲜活。我们不再在阴影下静立不动，开始漫步，也许会清神爽意或者在阴影之外走身向阳光。\n 漫步是什么？漫步可以是诗人在风景中的游走；漫步可以是人们阅读时在作者的世界中观赏；更可以是一种自心中的运动，换一个角度去看景色，就是阴影中阳光。\n 我们漫步阳光下，如果我们流着泪水，也许用不了多久心情就会平复；我们待在阴影中，不再流泪那么心情不过只保持着怒气与悲伤。为什么？流泪不是懦弱的象征吗？\n 泪水是什么？泪水可能只是一种水；泪水也许是一场磨难的产物；泪水也可以是一种情感的渲泄，一种平复心情的良药。\n 漫步在阳光下，流着泪，一定是喜极而泣了吗，乐景可平哀情，更可以助涨乐情，愿我们都能有一天，能漫步在阳光下，喜极而泣。 ——《阳光之下》	1. 文章中连接词使用不当的句子用橙色高亮部分标出；其中使用不当的连接词用蓝色高亮部分标出； 2. 以本篇文章为例，与“但”对应的连接词为“虽然”，文章中使用“更”这个连接词，视作连接词使用不当。
分句不合理	主要表现有两个： 1. 当断不断，一逗到底； 2. 不当断却断，割裂句子。	生活中有很多形形色色的人，有很多人值得我们敬佩，但这个人让我十分敬佩。\n 在长乐路的十字路口上，一直伫立着一位胖胖的交警。他不管春夏秋冬都穿着他的“荧光战衣”。他为人宽厚老实，就是有点太尊守规矩了，只要有人行驶不规范，他不管那个人有没有事都要接受到相应的惩罚。所以人们高兴时叫他“胖哥”，不高兴到时候叫他“肥猪”。\n 有一次我上学要迟到了。闷着头硬闯红灯。就在这时，一直粗糙的大手把我拉了回来，我回头看，“胖哥”正怒视冲冲地瞪着我，我刚一开口就被“胖哥”打断了。他严肃地对我说：“同学，你不能闯红灯，你不知道有多危险啊！你可知道这是十字路口啊！就你这样斜着穿	1. 文章中分句不合理的句子用绿色部分标出： 如果是属于第一种“当断不断，一逗到底”的情况，使用错误的标点符号

		过去，出了事故你付得起吗“……就这样，“胖哥”教育了二十多分钟，我成功的迟到了。心里正不断咒骂这“胖哥”：这个死肥猪，我闯红灯管你什么事？我是怎么你了，这么跟我过不去，他就是故意想让我迟到的！此后我就一直叫他“肥猪”。\n 过年从爷爷家回家的时候，十字路口的红绿灯坏了，十字路口的车都被各路车堵着，一时间，整个长乐路的汽笛声爆炸似地响着。这下可好了，我们都要在车上守岁了！过了一会儿，有一个荧光闪闪的巨人走了过来。这一看！啊！“胖哥”来救我们了！我们在“胖哥”的指挥下，不到十分钟就能正常通行了，路过“胖哥”身旁时，我看到了一个高大的身影，他的手正在不停地指挥，哨子紧紧地扣在嘴上，眼睛里充着血丝，但是炯炯有神。这让我对他不犹地产生了无以的敬意。\n“胖哥”没有回家过年，没有朋友陪着，但他活得很精彩，就因为这件事他一战成名，被分配到管理部门了。我虽再没见到他，但他一直是我最敬佩的人。他虽然是一个很小的人物，但是在我的眼里，在我群众眼里，他就是一位伟大的神。 ——《这样的人让我敬佩》	号用灰色高亮部分标出； 如果是属于第二种“不当断却断，割裂句子”的情况，使用错误的标点符号用紫色部分标出。
逻辑不通顺	文章上下文间逻辑关系使用不当	对人们来说，做事情，要做自己感兴趣的。什么是该做的，什么是不该做。在做事这一方面，每个人都是有人们自己的人生经历。\n 在人们的人生经历中，从小到大，做事是从我做起的。从出生至现在，人们小时候不懂事，非常喜欢乱做事，喜欢贪玩、捣蛋等，到了幼儿时期，比三岁更懂事，喜欢读书。到了六岁，读书的次数比之前越来越多，贪玩的次数越来越少。进入小学阶段，人们就开始学汉语拼音，开始学汉字。在人生的成长经历中，既有成功又有挫折。人们培养的好习惯越来越多，不好的越来越少。到了初中阶段，读书的次数越来越多。当初中成绩去考不好时，那就上不了高中，当高中成绩考不好时，就上不了大学。考完大学之后，会走向社会且找好工作，等等......\n 每个人做的事的结果是不一样的。人们做的事都要有兴趣是不一定正确的，对人们来说，人们每个能力有大有小，但是这点人都会有精神的，对于做事方面，有些事情是对人们感兴趣的。读书是对人们最感兴趣的，会让人们放松下来。\n 你从出生至现在，小时候是不肯读书且爱贪玩，长大之后，比之前懂事了，书就开始读的越来越多了。到了小学高年级阶段，书就读得越来越多了。到了初中阶段，书就读的越来越多，就越来越兴趣了。未来，到上高中、大学时期。书是要反	1. 文章中逻辑不通顺的上文和下文部分分别用深绿和深紫色高亮部分标出； 2. 以本篇文章为例，当作者写到“书是要反复读且要理解含义，就更会有兴趣了”，后面应该接和前文有解释说明关系的句子，而文中示例前后文间为转折关系。

		复读且要理解含义，就更会有兴趣了，下次要少玩电子产品中的游戏，用上网搜资料，这是不必要的。\n 做事是有时间的，大部分都不是一气呵成的。要有“少年强，则国强”的人生道路。 ——《对人们做的事》	

2.1.3 Expected Outputs

The submission should consist of a Python-based model, a development report that includes instructions for model usage, and prediction results of the testing datasets. It is crucial to ensure that the format of the model input is consistent with that of the testing datasets. To submit your work, please write the prediction results into a JSON file using the following format: **[{"id":(str), "sentence_quality":(dict)}，{…}, …]**, with the same sample order as the testing datasets.

Please note that:

- Improper use of conjunctions, mark as 1;
- Lack of logical flow between contexts, mark as 2 (if this error occurs, both sentences involved in the lack of logical flow are marked as 2);
- Illogical sentence breaks, mark as 3 (if this error occurs, both sentences involved in the illogical sentence breaks are marked as 3);
- Off-topic, mark as 4;
- No logical errors, mark as 0.

Below is a sample input and output for your reference.

Input Sample:

```
JSON
{
    "id": "468",
    "title": "夏日晚风自宜人",
    "text": [
            "那时一个相生共荣的小院，就在校篮球场旁。″,
           ″历经一节班会课”开不开电风扇“的争吵，自习课上终于回归了平
静。同桌轻轻拍了拍我说：”看黑板上写了什么？“我顺着看去，薄暮黄昏，散在
黑板上，墨绿的，托出一行精致的白色粉笔字”夏日晚风自宜人，不妨出去走走
“这是什么意思？一个高个的男生突然大嚷了一声”这是让我们把电风扇开到最大
档吧！“无人回应。这时，同桌邀我出去走走，我欣然应允。″,
           ″我们沿着沿着雪白的跑道，向前奔去。薄暮，暖黄的光穿过林立的教
学楼，透过高大的篮球架，洒在了小院里，不伤花谢，不羡柳青，花柳为木，树生
盎然，青叶在树梢上摇动，光影带来了最朴素真纯的生命风度，叶影婆娑间，绿起
```

```
人间四月天。日暮落在那一小丛月季上，显得浓烈而又庄重，月季的影子被一旁的栅栏轻轻牵住，不时微微晃动。″,
                ″” 好美“同桌惊讶的指着那树那花，他张开双臂，发丝被那日光携着向身后飘去。″,
                ″那是什么，我问着自己，是柳暗花明，是惊人月季？不，是风；风引导着我们与自然，与世间万物交融。″,
                ″在多少个日暮黄昏，我们倚在栏杆上说说笑笑，却意识不到那抹清凉；多少个日子，我们漫步在操场，金灿灿的枇杷果明明如耀眼宝万般晃动，我们却不自知。″,
                ″” 夏日晚风凉，少年亦如斯“。″,
                ″小时候都渴望成为一棵树，长大才明白，人不能成为树，不是因为不能像树一样高大，而是缺失树的干净、坚守、温暖的灵魂。风创造了千奇百怪的大自然，铸就了一棵棵独一无二的树。树，都能发现并体现大自然的美，人却难以做到。″,
                ″世间紧迫地需要一双发现美的眼睛，美，就在身边，就在大自然。″,
                ″倾听草木的呼唤，学着做一棵向着阳光的树。″
            ],
            "sentences": [
                 ″那时一个相生共荣的小院，就在校篮球场旁。″,
                ″历经一节班会课”开不开电风扇“的争吵，自习课上终于回归了平静。同桌轻轻拍了拍我说：”看黑板上写了什么？″,
                ″“我顺着看去，薄暮黄昏，散在黑板上，墨绿的，托出一行精致的白色粉笔字”夏日晚风自宜人，不妨出去走走“这是什么意思？一个高个的男生突然大嚷了一声”这是让我们把电风扇开到最大档吧！″,
                ″“无人回应。这时，同桌邀我出去走走，我欣然应允。″,
                ″我们沿着沿着雪白的跑道，向前奔去。″,
                ″薄暮，暖黄的光穿过林立的教学楼，透过高大的篮球架，洒在了小院里，不伤花谢，不羡柳青，花柳为木，树生盎然，青叶在树梢上摇动，光影带来了最朴素真纯的生命风度，叶影婆娑间，绿起人间四月天。″,
                ″日暮落在那一小丛月季上，显得浓烈而又庄重，月季的影子被一旁的栅栏轻轻牵住，不时微微晃动。″,
                ″” 好美“同桌惊讶的指着那树那花，他张开双臂，发丝被那日光携着向身后飘去。″,
                ″那是什么，我问着自己，是柳暗花明，是惊人月季？″,
                ″不，是风；风引导着我们与自然，与世间万物交融。″,
                ″在多少个日暮黄昏，我们倚在栏杆上说说笑笑，却意识不到那抹清凉；多少个日子，我们漫步在操场，金灿灿的枇杷果明明如耀眼宝万般晃动，我们却不自知。″,
                ″” 夏日晚风凉，少年亦如斯“。″,
            ″小时候都渴望成为一棵树，长大才明白，人不能成为树，不是因为不能像树一样高大，而是缺失树的干净、坚守、温暖的灵魂。″,
            ″风创造了千奇百怪的大自然，铸就了一棵棵独一无二的树。″,
            ″树，都能发现并体现大自然的美，人却难以做到。″,
            ″世间紧迫地需要一双发现美的眼睛，美，就在身边，就在大自然。″,
            ″倾听草木的呼唤，学着做一棵向着阳光的树。″
        ]
}
```

Output Sample:

```JSON
{
    "id": "468",
    "sentence_quality": {
            "那时一个相生共荣的小院，就在校篮球场旁。": [
                2,
                4
            ],
            "历经一节班会课”开不开电风扇“的争吵，自习课上终于回归了平静。同桌轻轻拍了拍我说：”看黑板上写了什么？": [
                2
            ],
            "“我顺着看去，薄暮黄昏，散在黑板上，墨绿的，托出一行精致的白色粉笔字”夏日晚风自宜人，不妨出去走走“这是什么意思？一个高个的男生突然大嚷了一声”这是让我们把电风扇开到最大档吧！": [
                2
            ],
            "“无人回应。这时，同桌邀我出去走走，我欣然应允。": [
                2
            ],
            "我们沿着沿着雪白的跑道，向前奔去。": [
                0
            ],
            "薄暮，暖黄的光穿过林立的教学楼，透过高大的篮球架，洒在了小院里，不伤花谢，不羡柳青，花柳为木，树生盎然，青叶在树梢上摇动，光影带来了最朴素真纯的生命风度，叶影婆娑间，绿起人间四月天。": [
                0
            ],
            "日暮落在那一小丛月季上，显得浓烈而又庄重，月季的影子被一旁的栅栏轻轻牵住，不时微微晃动。": [
                0
            ],
            "”好美“同桌惊讶的指着那树那花，他张开双臂，发丝被那日光携着向身后飘去。": [
                0
            ],
            "那是什么，我问着自己，是柳暗花明，是惊人月季？": [
                0
            ],
            "不，是风；风引导着我们与自然，与世间万物交融。": [
                0
            ],
            "在多少个日暮黄昏，我们倚在栏杆上说说笑笑，却意识不到那抹清凉；多少个日子，我们漫步在操场，金灿灿的枇杷果明明如耀眼宝万般晃动，我们却不自知。": [
                2
            ],
            "”夏日晚风凉，少年亦如斯“。": [
                2
            ],
            "小时候都渴望成为一棵树，长大才明白，人不能成为树，不是因为不能像树一样高大，而是缺失树的干净、坚守、温暖的灵魂。": [
                0
            ],
            "风创造了千奇百怪的大自然，铸就了一棵棵独一无二的树。": [
                0
            ],
            "树，都能发现并体现大自然的美，人却难以做到。": [
                0
            ],
            "世间紧迫地需要一双发现美的眼睛，美，就在身边，就在大自然。": [
                0
            ],
            "倾听草木的呼唤，学着做一棵向着阳光的树。": [
                0
            ]
        }
}
```

2.1.4 Training Datasets

We offer approximately 500 Chinese essays written by middle school students, of which 400 can serve as training sets and 100 as verification sets. Each data sample includes the title, the body of the article and the quality of each sentence within the text. Participants are also welcome to utilize data from other sources, such as manual annotation or automatic annotation using models or tools, to enhance their training experience.

2.1.5 Testing Datasets

We offer a comprehensive collection of 5,000 Chinese essays that serve as our testing datasets. These valuable resources are made available to participants in the form of a JSON file that includes key information, such as the essay's ID, title, and text content in the format of **[{"id" : "", "title" : "" , "text" : [] , "sentences" : []} ...]**.

To ensure the highest standards of accuracy and quality, we meticulously select a portion of the data in the test set for review. This enables us to provide insightful feedback to participants and further refine their method.

2.1.6 Evaluation Metrics

The evaluation of the discourse logic error detection in this task will use precision (P), recall (R), and macro-F1 score (F1-score, F1). Precision is calculated as the number of correctly identified logical errors in sentences divided by the total number of identified logical errors in sentences. Recall is calculated as the number of correctly identified logical errors in sentences divided by the total number of annotated logical errors in sentences. F1-score is calculated as (2*P *R)/(P +R).

2.2 Topic Coherence Modeling (TCM)

2.2.1 Task Description

The foundational basis of an article's coherence lies in the logical layout of the content and the clarity of the central ideas. In writing, regardless of whether one uses Chinese, English or another language, the key is to develop around a core topic. Each paragraph typically comprises a topic sentence that encapsulates the central idea of the paragraph, and several supporting sentences, which serve to explain, describe, or prove the topic sentence. The topic sentence plays a crucial role in highlighting the paragraph's central thought and guiding the development of the paragraph. Therefore, modeling the coherence of the topic sentence is essential for evaluating the overall coherence of the entire article, which is referred to as Topic Coherence Modeling.

2.2.2 Task Definition

Given an essay by a middle school student, the task includes identifying the topic sentence of each paragraph and evaluating the logical coherence between the topic sentences. The topic sentence is a complete sentence that summarizes, narrates, or explains the topic of the paragraph and may be located at the beginning, middle, or end of the paragraph. In some cases, a paragraph may not contain an explicit topic sentence but conveys it implicitly through the entire content. Annotators need to identify the topic sentence of each paragraph and determine which sentence best represents the central idea of the entire text (the thesis sentence). Subsequently, assess the logical relationship between adjacent topic sentences, analyzing how they interconnect to support the overall structure of the article. The definition and examples of logical relations are as follows:

粗粒度关系	细粒度关系	定义	例句	
			句子 1	句子 2
共现关系 Co-occurrence Relationship	并列关系 Parallel Relationship	描述同一事件的几个方面、相关的几件事情或相对的情况，在意义上并存、共现或对立，在语序上可以调换顺序且不改变句义。	我的老爸就像一只鸡，每天都很早起，然后上班，每次他都很早起，早饭就吃两口，再上班，而且很晚睡觉。	我就像一个变色龙，总是变脸，我有时不开心的时候就板着脸，我一开心就一直微笑，我伤心的时候就外表，我的衣服也会变哦。
	顺承关系 Sequential Relationship	篇章单位之间存在时间、空间、步骤、逻辑事理上的先后顺序，包括顺序和逆序两种情况；但不包括同时发生的事件，同时的事件属于并列关系。由于存在先后关系，篇章单位的顺序不可随意调换。篇章单位的主体可以是同一个人或事物，也可以是不同的人或事物。	我看见女娲先杀死了一只大乌龟，用它的腿撑着天空。	接着，杀死了一只黑龙。
	递进关系 Progressive Relationship	后一篇章单位在数量、质量、范围、时间等方面比前一篇章单位更进一层，强调程度的增强加深；篇章单位的顺序通常不可调换。与顺承相比，顺承只表现为一种先后顺序，没有程度的加深；而递进关系则强调后者比前者在程度上更进	图书馆里有各种各样的图书，种类数也数不清。	甚至连英文书都有呢。

		一步。		
	对比关系 Contrastive Relationship	对比关系是指在文本中出现的两个或多个事物、概念、观点、行为、状态等之间的明显的、直接的、相对的差异或相似性。对比关系通常用于强调、比较和对照。	他的衣服很破旧，袖口打了几个补丁，颜色也有些发白。	但是却洗得很干净，没有一点油污。
过渡关系 Transition Relationship	让步关系 Concessive Relationship	某一篇章单位提出一个假设的事实，并且退让一步暂且承认这个假设的真实性，另一篇章单位叙述一个与之相反或相对的情况。语序上，假设的事实通常在前。	她虽然不用功学习。	考试却及格了。
	转折关系 Turn Relationship	某一篇章单位提出一个客观事实，另一篇章单位叙述一个与之相反或相对的情况。语序上，转折部分一般在后，有时也会倒装变化。	月亮发出的黄色光芒，把周围的几朵灰灰的云也照黄了。	但又仔细一看，好像月亮也不是纯黄色的，有黑乎乎的东西在上面。
解说关系 Explanatory Relationship	泛化关系 Generalization Relationship	后一篇章单位是对前一篇章单位的概括、总结和泛化；篇章单位间不可调换顺序，否则转变为细化关系。与之相关的连接成分包括“总之”、“总体而言”、“综上所述”等；与细化类似，泛化关系也主要依靠意义的制约。	为了达到这个目的，他们讲究亭台轩榭的布局，讲究假山池沼的配合，讲究花草树木的映衬，讲究近景远景的层次。	总之，一切都要为构成完美的图画而存在，决不容许有欠美伤美的败笔。
	细化关系 Specification Relationship	后一篇章单位是对前一篇章单位的细化描述，包括举例、解释、说明、补充等；篇章单位间不可调换顺序，否则转变为泛化关系。与之相关的连接成分包括“这”、“即”、“例如”、“也就是说”等；但与其他类型相比，细化关系在多数情况下没有提示成分，通常表现为词义或句义的关联。	我的老妈就像一个母老虎，我一不听话她就发脾气。	有一次，我没有写完作业她就发脾气，说："你怎么还没写完啊！。"

主从关系 Dominant-Subordinate Relationship	客观因果关系 Objective Causal Relationship	某一篇章单位说明原因，另一篇章单位说明由该原因导致的结果，两者均是客观事实。原因和结果的前后位置不固定，但二者有主次之分，有时原因为主，有时结果为主，视句义而定。	我查了书籍，原来农历十五、十六都为满月。	所以今天的月亮也是最大最圆的。
	背景关系 Background Relationship	篇章中经常出现事件、地点、历史等情况的介绍，此类环境信息与篇章正文构成背景关系。在背景关系中，某一篇章单位交代事情发生的历史情况、现实环境、前情概要等，如时间、地点、历史背景、政治环境等，另一篇章单位叙述事情的内容。背景关系具有特定的限制条件：如果事件背景和内容之间发生因果、转折等其他主从关系，则优先标注其他关系；只有单纯的环境描写才算作背景关系。语序上，背景部分通常在事件内容之前。	迪士尼乐园是人们向往的地方，也是周末玩耍的好去处。	今天我就给大家推荐上海迪士尼乐园。
	特定条件关系 Specific Conditional Relationship	某一篇章单位提出特定的条件，另一篇章单位说明以该条件为依据推断出的结果。其中，特定条件可以包括充足条件，代表的格式是“只要……就……”；可以是必要条件，常用的连接成分有“只有……才……”、“除非……否则……”等；也可以是周遍性条件，常使用“无论……都……”、“不管……也……”等格式。	只有坚持锻炼。	才会有好身体。
	假设条件关系 Hypothetical Conditional Relationship	某一篇章单位提出虚拟性条件，另一篇章单位说明该假设条件实现后所产生的结果，或为了实现该假设条件而应采取的措施。	如果我们好好学习。	就能取得好成绩。

	主观推论关系 Subjective Inference Relationship	某一篇章单位说明事实依据，另一篇章单位说明由此推断出的主观结论；与客观因果关系所不同的是，推论得到的结果是主观的。语序上，事实依据往往在前，主观结论在后；二者有主次之分，结论通常是句义的核心。	去之前一定要提前预约！	不然你可能会排两个小时的队！

2.2.3 Expected Outputs

To submit your work, please ensure that it includes a Python-based model, a development report detailing the instructions for using the model, and the prediction results for the testing datasets. It is essential that the input format of the model matches the format of the testing datasets. When preparing your submission, write the prediction results into a JSON file with the following structure: [{"id": "(str)","paraTopicLogicRelation": [{"Arg1": {"paragraphIdx": "(int)", "ParagraphTopic": "(str)"},"Arg2": {"paragraphIdx": "(int)","ParagraphTopic": "(str)"}, "Relation": "(str)"}, …] },...]. The “paraTopicLogicRelation” section should be formatted as follows: Each entry contains two arguments, “Arg1” and “Arg2”, each represented as a dictionary with “paragraphIdx” and “ParagraphTopic” as keys. Additionally, the “Relation” key indicates the relationship between the two arguments as a string. Ensure that the sample order in the JSON file matches that of the testing datasets.

Below is a sample input and output for your reference.

- Input Sample:

```
JSON
{
    "id": "3027",
    "title": "学会“读”",
    "text": [
        "在生活中，人们把学习的人叫学者，把研究艺术人叫艺术爱好者，把做演讲演说多人叫演讲者。那么，一个读书人叫什么，没错，是读者。",
        "其实“读者”是对读书人的一种赞誉，正因为这一点，所以读书之人不一定就能成为一个合格的读者。要学会读书，才是一个读者。",
        "读者，要学会读背景。一个真正喜欢读书人，是不会仅仅把全书看完一遍就了事了的。读书前，先把书的背景了解，可能读时会更能理解书中想表达的意思，比如说《儒林外史》一书，如果不了解作者当时所处的社会环境是那么地腐败
黑暗，你会把这本书当一本好笑小说。确定，《儒林外夫》中吴敬梓的言辞十分白话，情节内容真的再有趣不过了，但这也主是吴敬梓想要达到的，他想用这些过分荒淡、好笑到人和事，反映出当时的社会是那么的可笑而更可悲啊！了解了书的背景,才能真正体现到书中的那种讽刺与作者的无奈。",
            "读者，要学会思考。《论语》中道：“学而不思则罔”，读书也是一种学习的过程，同样也需要读者对书进行思考和探究，很多书内容有些声，可能有时你还会对作者的观点有所不理解，甚至否认。但如果你在了解书背景的同时结合书中的内容加以深究，就会有不同的感受。就像《朝花夕拾》中鲁迅在《王倡会》的描绘的父亲形象，是那么的严厉，让人觉得鲁迅是在对自己的父亲表示讨厌，对父亲十分不喜爱，但你再以鲁迅所生的环境，想一想，你就会顿开茅塞，鲁迅这里并不是在怪自己的父亲，而是想通过这件事，表现出旧中国封建的思想教育方式抹杀了为孩子的天性。正就是思考的好处。",
            "读者，更要品读。品读也可以说是复读。很多的人会对一些名著进行复读，品析内容。其实复读更有利于让对书产生理解与共鸣，古人言：“读书百遍，奇异自现”也正如此，每一遍读你都会有新的感悟。",
            "其实这三种方式读书也同样可以运用于生活，生活中也需要这样认真的态度以面对。要学会做一个读者，也更要有才为读者后学习书中之道，做一个生活的享受者。"
    ]
}
```

- Output Sample:

```
JSON
{
    "id":"3027",
    "paraTopicLogicRelation": [
            {
                "Arg1": {
                    "paragraphIdx": 0,
                    "ParagraphTopic":"那么，一个读书人叫什么，没错，是读者。"
                },
                "Arg2": {
                    "paragraphIdx": 1,
                    "ParagraphTopic":"要学会读书，才是一个读者。"
                },
                "Relation": "解说关系"
            },
          ...
]
 }
```

2.2.4 Training Datasets

We offer approximately 452 Chinese essays written by middle school students, of which 400 can serve as training sets and 52 as verification sets. Each data sample includes the title, the body of the article and the logical relation of each paragraph. Participants are also welcome to utilize data from other sources, such as manual annotation or automatic annotation using models or tools, to enhance their training experience.

2.2.5 Testing Datasets

We offer a comprehensive collection of 5,000 Chinese essays that serve as our testing datasets. These valuable resources are made available to participants in the form of a JSON file that includes key information, such as the essay's ID, title, and text content in the format of **[{"id": "", "title": "","text": []} ...]**.

To ensure the highest standards of accuracy and quality, we meticulously select a portion of the data in the test set for review. This enables us to provide insightful feedback to participants and further refine their method.

2.2.6 Evaluation Metrics

In evaluating paragraph topic sentence accuracy, we adopt the Topic Sentence Accuracy (TSAcc) metric, which requires an exact match for recognition of the correct topic sentences within a paragraph. Additionally, we may use more lenient metrics for assessment such as BLEU or BERTScore. For evaluating the relationship between topic sentences, we employ precision (P), recall (R), macro-F1 score (F1-score, F1) and accuracy (Acc), and other relevant metrics.

In terms of precision and recall, the methodology follows that of the assessment of recognition performance for logical relationships between paragraphs. Precision is defined as the ratio of the number of correctly identified instances over the total number of instances flagged by the system. Recall is the proportion of correctly identified instances in relation to the total instances present in the reference data. The macro-average F1 score (F1), which combines precision and recall, is computed as the harmonic mean of the two: $F1 = 2 * (P * R) / (P + R)$.

2.3 Discourse Coherence Feedback Generation (DCFG)

2.3.1 Task Description

Although large language models (LLMs) have become increasingly widespread in text processing tasks, their use in assessing and generating feedback on discourse coherence is still relatively nascent. Current research primarily focuses on the accuracy of the evaluation, with little attention given to providing substantive recommendations for improvement based on the evaluation results. The objective of this task is to explore how to combine LLM technology to provide concrete, actionable suggestions for the revision of compositions, thereby offering more valuable support for writing instruction. This requires not only an in-depth analysis of the discourse coherence of compositions but also the ability to generate instructive feedback based on the analysis results, which is referred to as Discourse Coherence Feedback Generation.

2.3.2 Task Definition

Given a middle school student's essay, the annotator will provide feedback based on the coherence of the discourse.

This includes four aspects.

1. Discourse Analysis: understanding the overall structure and fluency.
2. Topic Extraction: identifying the core argument and supporting details.
3. Logic Detection: evaluating the logical sequence and relevance of the arguments.
4. Coherence Rating: giving a comprehensive rating based on the above analysis.

The comments should be specific and constructive, guiding students on how to improve their writing.

2.3.3 Expected Outputs

The submission should consist of a Python-based model, a development report that includes instructions for model usage, and prediction results of the testing datasets. It is crucial to ensure that the format of the model input is consistent with that of the testing datasets. To submit your work, please write the prediction results into a JSON file using the following format: **[{"id":(str), " comment ":(str), ...]**, with the same sample order as the testing datasets.

Below is a sample input and output for your reference.

- Input Sample:

```
JSON
{
    "id": "1326",
    "title": "仙人掌",
    "text": [
        "生活中，随处可见的植物，都有一个不凡的精神。陶渊明爱的菊，周敦颐爱的莲，都有高洁的精神。这些精神，往往会令人感到受益匪浅。",
        "我们家原来有一盆仙人掌，长得很丑陋。仙人掌像几个绿色的手掌，有大有小，有粗有细。这几个“手掌”上还长满了几个小包，小包上面长满了刺。只要人一不小心，碰到了，就会扎到皮里。更令人讨厌的是，这种刺特别小，要照着光才能挑下来。这个又丑又烦的仙人掌，每年都能开花。因为这个开的金黄色大花，我们才不舍得扔它。而在它旁边的另一盆小野花，虽然没有名字，但是那些日出开、日落合的花儿，淡紫色的花瓣映衬着阳光，颜色又格外的鲜艳。不多说，与那仙人掌相比，这可是美丽了不知多少倍。",
        "在某一年的暑假，我们一家去外地旅游，我们出游前忘记把阳台的植物搬回来了，再加上南京那干热的气候，像大火球一般的太阳，恐怕这两个植物都受不了吧！",
        "我们回到家，去阳台看望了两个可怜的植物。小野花因为没有水的摄入和太阳的暴晒，盆里的土干得有了一些细细的裂纹，花朵垂了下来，叶子也有微微泛黄；而仙人掌呢，他并没有被这炎热的天气晒蔫掉，反而好像比之前更绿了一些。",
        "仙人掌是生活在南北回归线之间的沙漠里。那里位于热带，雨水稀少，空气干燥，而仙人掌却是十分的耐旱。",
        "在我们的生活中，也有生活让自己不如意的地方。但我们就像仙人掌一样，用自己的实力，与这样的挫折做斗争，磨炼自己不服输的意志。也许你在不被人认可时，你也可以通做自己的努力来证明自己。"
    ]
}
```

- Output Sample:

```
JSON
{

    "id": "1326",

    "comment": "本篇作文《仙人掌》通过对比仙人掌与其他植物的生存特性，巧妙地阐述了在逆境中坚韧不拔的生命力与自我证明的重要性。文章开头通过引用陶渊明对菊花的喜爱和周敦颐对莲花的赞美，巧妙地引入了仙人掌这一主题，展现了不同植物所代表的精神价值。\n\n 随后，通过生动描绘家中仙人掌的丑陋外表与其生命力的对比，作者引出了仙人掌即使在恶劣环境下
```

```
        依然能够绽放美丽花朵的独特魅力。这种鲜明的对比，不仅让读者对仙人掌
        这一植物有了更深的认识，也寓意了即使在困难和不被理解的情况下，依然
        可以通过自身的努力展现出独特的价值。\n\n 文章通过实例讲述了一次假期
        旅行中，仙人掌和小野花在极端环境下的生存状态，进一步强调了仙人掌顽
        强生存的能力。最后，作者将仙人掌的特性与人生的挑战相比拟，鼓励人们
        在面对困难时要有仙人掌一样的韧性和自我证明的勇气。\n\n 总体而言，文
        章逻辑性强，主题明确，通过具体的例子和生动的描写，成功传达了坚韧不
        拔和自我证明的主题。语篇连贯度评级为优秀，体现了作者较高的思维能力
        和语言表达能力。"

}
```

2.3.4 Training Datasets

We offer approximately 450 Chinese essays written by middle school students, of which 400 can serve as training sets and 50 as verification sets. Each data sample includes the title, the body of the article and the composition feedback. Participants are also welcome to utilize data from other sources, such as manual annotation or automatic annotation using models or tools, to enhance their training experience.

2.3.5 Testing Datasets

We offer a comprehensive collection of 5,000 Chinese essays that serve as our testing datasets. These valuable resources are made available to participants in the form of a JSON file that includes key information, such as the essay's ID, title, and text content in the format of **[{"id":"","title":"","text":[]} ...]**.

To ensure the highest standards of accuracy and quality, we meticulously select a portion of the data in the test set for review. This enables us to provide insightful feedback to participants and further refine their method.

2.3.6 Evaluation Metrics

In the discourse coherence feedback generation task, evaluation metrics comprise PPL, BLEU, ROUGE-L, BERTScore, and Human Expert Assessment. PPL measures feedback fluency (lower is better); BLEU assesses n-gram overlap with expert comments (higher indicates better alignment). ROUGE-L gauges recall of key coherence points (higher means better coverage). BERTScore evaluates token-level semantic similarity (higher signifies enhanced coherence detection). Human evaluation qualitatively judges feedback relevance, actionability, clarity, and overall coherence.

Given the multifaceted nature of discourse coherence feedback, a weighted combination of these metrics can be employed to derive a comprehensive evaluation score. The possible weight allocation is as follows: Score = 15% * PPL + 25% * BLEU + 25% * ROUGE-L + 30% * BERTScore + 5% * Human Expert Assessment.