# Plot Retrieval as an Assessment of Abstract Semantic Association

**Shicheng Xu**[*1,2] **Liang Pang**[1†] **Jiangnan Li**[2] **Mo Yu**[2†]
**Fandong Meng**[2] **Huawei Shen**[1] **Xueqi Cheng**[1] **Jie Zhou**[2]
[1]CAS Key Laboratory of AI Security, Institute of Computing Technology, CAS
[2]Pattern Recognition Center, WeChat AI
xushicheng21s@ict.ac.cn pangliang@ict.ac.cn moyumyu@global.tencent.com

## Abstract

Retrieving relevant plots from the book for a query is a critical task, which can improve the reading experience and efficiency of readers. Readers usually only give an abstract and vague description as the query based on their own understanding, summaries, or speculations of the plot, which requires the retrieval model to have a strong ability to estimate the abstract semantic associations between the query and candidate plots. However, existing information retrieval (IR) datasets cannot reflect this ability well. In this paper, we propose PLOTRETRIEVAL , a labeled dataset to train and evaluate the performance of IR models on the novel task *Plot Retrieval*. Text pairs in PLOTRETRIEVAL have less word overlap and more abstract semantic association, which can reflect the ability of the IR models to estimate the abstract semantic association, rather than just traditional lexical or semantic matching. Extensive experiments across various lexical retrieval, sparse retrieval, dense retrieval, and cross-encoder methods compared with human studies on PLOTRETRIEVAL show current IR models still struggle in capturing abstract semantic association between texts. PLOTRETRIEVAL can be the benchmark for further research on the semantic association modeling ability of IR models.

## 1 Introduction

We propose a new task, *Plot Retrieval*, which retrieves the relevant plots from the book for a query. The task is a spontaneous process in humans' daily lives. When reading a book or coming across other life events that remind a plot, humans naturally require to find the target plot. As a result, *Plot Retrieval* is a common and natural scenario but has not been well-studied in NLP.

Although *Plot Retrieval* can be formalized as an information retrieval (IR) task, the key challenge in *Plot Retrieval* is estimating the abstract semantic association between two texts that cannot be simply measured by lexical or semantic matching. Specifically, we analyze the logs of online reading apps such as Kindle, iReader, Douban[1] and find that the semantic association between the description of the plot given by the reader (i.e., query) and the actual plot in the book is very abstract. This abstract association is mainly because users integrate their own understanding, summaries, or speculations of the plot when writing the query, which makes it hard to directly associate plots to the query like traditional lexical matching, semantic similarity, or relevance. For example, for a plot: *The stranger said "Monsieur Baron, there is a thief and murderer in your house." Baron shook and said "In my house?" The stranger, calm and collected, dropped his cap as he stretched his arms, and continued: "Monsieur Baron, watch out for murderers and thieves"*, and the query for this plot given by reader is *There is a person saving the Baron.* This association is generated by human

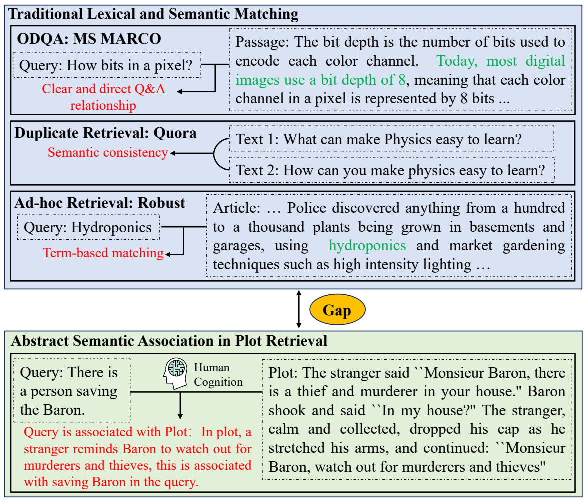


Figure 1: The gap between existing IR datasets and *Plot Retrieval* in estimating the relationship between texts.

*Work done during the Tencent Rhino-bird Research Elite Program at WeChat.
†Corresponding authors.

[1]https://www.ireader.com.cn, https://book.douban.com.



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 146–161
August 11-16, 2024 ©2024 Association for Computational Linguistics

cognition and is more difficult to estimate than just lexical or semantic matching because it requires IR models to understand that *the stranger reminds Baron to watch out for murderers and thieves* is actually associated with *saving the Baron*, even though their literal meanings are different. However, as shown in Figure 1, existing IR datasets do not reflect this abstract semantic association well. For example, in Open-domain Question-Answering such as MS MARCO (Nguyen et al., 2016), Natural Questions (Kwiatkowski et al., 2019), and SQuAD (Rajpurkar et al., 2016), the query and its corresponding passage form a clear and direct question-and-answer relationship. In Duplicate Retrieval such as Quora and MRPC (Dolan et al., 2004), the annotation is based on whether the semantics of the two texts are consistent. In Ad-hoc Retrieval such as Robust04 (Voorhees, 2004), lexical matching still accounts for the main part and semantic association is less (Xu et al., 2022).

A dataset that can reflect abstract semantic associations between texts generated by human cognition is important for the entire IR community to study the upper limit of the IR models' ability to model semantic association. However, it is very difficult to obtain the annotated query-passage pairs with sufficient abstract semantic association. Annotating abstract semantic association pairs requires annotators to pay the high reading cost for passage, and have sufficient comprehension ability to write a query that looks very different from the passage but has abstract semantic association with it.

In this paper, for *Plot Retrieval*, a novel and challenging IR task, we propose a labeled dataset called PLOTRETRIEVAL with 430K query-plot pairs. Compared with existing IR datasets, text pairs in PLOTRETRIEVAL have the following obvious characteristics: (1) more abstract semantic association generated by human cognition and (2) less word overlap. These two characteristics enable PLOTRETRIEVAL not only to be used to perform training on *Plot Retrieval* task but also become the benchmark for evaluating the ability of IR models to estimate abstract semantic association between texts. In the construction of PLOTRETRIEVAL, we collect publicly available raw data from the Internet, which shares the idea with (Wan et al., 2019; Yu et al., 2023). To address the difficulty in annotation mentioned above, instead of directly asking the annotators to write a query that has abstract semantic association with the plot, we first use weakly supervised information to collect query-plot pairs that may have semantic association, and let the annotator select the pairs that really contain abstract semantic association, regularize these pairs, and get the final query-plot pairs.

In experiments, first, we evaluate various lexical retrieval, sparse retrieval, dense retrieval, and cross-encoder methods trained on mainstream IR datasets such as MS MARCO on PLOTRETRIEVAL, and find that these methods do not perform well, which shows the difference between PLOTRETRIEVAL and the current IR datasets. A noteworthy finding is that BM25, the strong zero-shot IR baseline based on lexical-matching (Thakur et al., 2021; Izacard et al., 2022), achieves better performance on BEIR (Thakur et al., 2021) than many neural IR models, but has worse performance on PLOTRETRIEVAL. This indicates that PLOTRETRIEVAL has the higher challenge for semantic understanding rather than simple literal matching. Second, we train IR models on our weakly supervised data and achieve better performance than the models trained on MS MARCO, which indicates the effectiveness of our annotation strategy. Third, human studies show that the current IR models are far behind human in capturing abstract semantic association, and there is a lot of room for improvement in future research. Our contributions are:

- We propose a novel, critical and challenging task called *Plot Retrieval*, design a novel evaluation metric called N-RODCG and construct a dataset called PLOTRETRIEVAL for this task.
- Extensive experiments across various IR models and the comparison with human studies on PLOTRETRIEVAL show that the current IR models still struggle in capturing abstract semantic association between texts and there is a lot of room for improvement in the future research.
- We broaden the research field of Information Retrieval from lexical or semantic matching to more ambiguous abstract semantic association between texts, and PLOTRETRIEVAL can be used as an effective benchmark for evaluating this ability of IR models. We will release both English and Chinese versions of PLOTRETRIEVAL at https://github.com/xsc1234/Plot-Retrieval for further research.

## 2 Related Work

**Information Retrieval Datasets** According to specific task, existing mainstream IR datasets can be divided into: *Open Domain Question-answering* (MS MARCO (Nguyen et al., 2016), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2016), WebQuestions (Berant et al., 2013), FiQA (Maia et al., 2018), HotPotQA (Yang et al., 2018a) and CuratedTREC (Baudis and Sedivý, 2015), etc.), *Ad-hoc Retrieval* (Robust (Voorhees, 2004), ClueWeb (Yang et al., 2018b), MQ2007 (Qin et al., 2010)), *Duplicate Retrieval* (Quora, CQADupStack (Hoogeveen et al., 2015), MRPC (Dolan et al., 2004)), *Entity Retrieval* (DBPedia-Entity (Hasibi et al., 2017)), *Argument Retrieval* (ArguAna (Wachsmuth et al., 2018) and Touchè-2020 (Bondarenko et al., 2020)), *Citation Prediction* (SCIDOC (Cohan et al., 2020)) and *Fact Checking* (FEVER (Thorne et al., 2018) and Climate-FEVER (Diggelmann et al., 2020)). Existing datasets also cover a range of different domains of target documents like *Bio-Medical articles* (Tsatsaronis et al.), *Tweets* (Suarez et al., 2018), *News* (Soboroff et al., 2019).

In all the above datasets, the matching between texts can be summarized as a combination of lexical and semantic matching. The relationship of query-passage pairs in these datasets can usually be judged only by the literal meaning, without the need to deeply understand the semantics and judge the abstract association between semantics. Direct evidence is that BM25 (Robertson et al., 1995) can significantly defeat many neural IR models that have been trained on large-scale supervised datasets only through lexical matching on these datasets in the zero-shot setting (Thakur et al., 2021). PLOTRETRIEVAL has more abstract semantic association and less word overlap between texts, which is a more challenging dataset for IR models.

**IR Datasets for Books** Our dataset also extends into the significant domain of narrative literature for IR applications. While there exists an extensive list of datasets on story understanding (for more details, please refer to the survey (Sang et al., 2022)), there has been limited work addressing the IR aspect within the context of stories. In relation to our work, two other datasets are noteworthy. The first is *RELiC* (Thai et al., 2022), which frames the task as utilizing literary analysis paragraphs to retrieve quoted text. This task essentially falls within the realm of IR, although it lacks a standard format of IR queries. The second is *NarrativeQA* (Kočiskỳ et al., 2018), primarily designed as a book QA dataset but adaptable for an IR task (Frermann, 2019; Mou et al., 2021). However, it comes with a limitation that it does not provide groundtruth for the retrieval purposes.

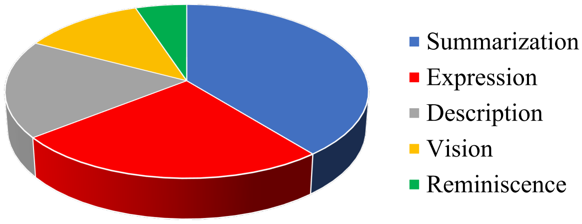


Figure 2: Statistics of abstract semantic association.

## 3 Task Description

### 3.1 Abstract Semantic Association

In the analysis of public data of online reading apps, we conclude five main manifestations of abstract semantic association between the query and the plot. (1) Query abstractly summarizes the plot (Summarization). (2) Query expresses feelings, analysis or comments about the characters or events in the plot (Expression). (3) Query depicts the characters in the plot (Description). (4) Query describes the overall visual information formed by the environment, characters, and events in the plot (Vision). (5) Query is motivated by the event in the plot to reminisce another related event (Reminiscence). Their statistics are shown in Figure 2.

### 3.2 Task Definition

*Plot Retrieval* aims to retrieve the relevant plots from the book for a query. Specifically, given a query $q$, a candidate set of plots $P = \{p_1, p_2, ..., p_n\}$ for a book and each plot $p_i$ consists of $m$ sentences ($m$ is a hyperparameter and we set it as 3). The model needs to give the ranking score for each $p_i \in P$ based on the association between plot $p_i$ and query $q$, rank the plots in $P$ according to the score, and return a list $R$ with Top-K plots. The challenge of this task is mainly in two aspects: (1) The semantic association between the query and the plot is very abstract. This is mainly because users integrate their own understanding, summaries, or speculations of the plot when writing descriptions. IR models struggle in identifying this abstract association. (2) Plots in the candidate set $P$ come from the same book, they have semantic and entity relatedness to each other. It makes IR models hard to distinguish the semantic difference.

### 3.3 Evaluation Metric: N-RODCG

As for the evaluation metrics for *Plot Retrieval*, in addition to the common information retrieval metrics, such as MRR (Mean Reciprocal Rank) and Recall, we propose N-RODCG (Normalized Relative Offset Distance Discounted Cumulative Gain), a novel metric that is more in line with the actual reading scene. The motivation of this metric is that each plot of the candidate set is actually the segment of continuous texts in the original book, even if the retrieved plot is not exactly the ground-truth plot, as long as it is close enough to the ground-truth plot in the original book, the ground-truth plot will appear in the reader's field of vision and be noticed by the reader. In addition, there is the strong semantic association between plots with small distances. N-RODCG comprehensively measures the ranking of the plots in $R$ and their distance from the ground-truth plots. For a query $q$, given a retrieved list of plots $R = [p'_1, p'_2, ..., p'_k]$ obtained from the model. Because each plot $p'_i$ consists of $m$ sentences, we can get the position of $p'_i$ in the original text of the book, which is the average value of each sentence index in $p'_i$ and we call it $s_i$. Then the positions of the plots in $R$ are $S = [s_1, s_2, ..., s_k]$. And the positions ($t_i$) of the ground-truth plots for $q$ is $T = [t_1, t_2, ..., t_g]$, $g$ is the number of ground-truth plots. The relative offset distance $d_i$ between $p'_i$ and ground-truth plots of $q$ can be computed as:

$$d_i = \min(|s_i - t_1|, |s_i - t_2|, ..., |s_i - t_g|). \quad (1)$$

Then, we define the Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002) between ROD and the ranking of the retrieved plots:

$$\text{RODCG@}k = \sum_{i=1}^{k} \frac{f(d_i)}{\log(i+1)}, \quad (2)$$

where $i$ is the ranking of plot $p'_i$, $f$ is the piecewise function ($\alpha$ is the window of the reader's field of vision and we set it to 5 based on statistical data):

$$f(d_i) = \begin{cases} \frac{1}{d_i+1}, & d_i < \alpha; \\ 0, & otherwise. \end{cases} \quad (3)$$

N-RODCG can be computed as:

$$\text{N-RODCG@}k = \frac{\text{RODCG@}k}{\text{I-RODCG@}k}, \quad (4)$$

I-RODCG is the value when the plots in retrieval list $R$ for $q$ are optimally ranked, that is, the theoretical maximum value of N-RODCG.

## 4 PLOTRETRIEVAL

We introduce collection, filtering, translation, annotation, and statistics for PLOTRETRIEVAL in this section. More details are introduced in Appendix.

### 4.1 Overview of Dataset Construction

The row data of PLOTRETRIEVAL is collected from an online reading app on the Internet. Specifically, we notice recent reading apps allow readers to write publicly available comments on the texts in the book. Many of these comments include abstract descriptions of the plots in the corresponding texts. They are written by the readers based on their own understanding during book reading. While they are semantically associated with the plots, they require sufficient comprehension ability to discover and are challenging for IR models to identify. These comment-plot pairs constitute the weakly supervised signal for query-plot pairs in PLOTRETRIEVAL. We first filter these pairs to remove the comments that have obvious word overlap with plots or have little practical meaning. However, the filtered comment-plot pairs still cannot be directly used as PLOTRETRIEVAL, because the comments written by readers are free-style and have a lot of noise. We let the annotators do more identification and rewriting on them. After the human annotation, we exploit the labeled datasets to construct an automatic annotation model for fast, low-cost acquisition of large datasets. Last but not least, we ensure the complete independence of the training set and the test set during the construction of PLOTRETRIEVAL, which makes that there are enough differences in the domain between the training set and the test set to more reasonably evaluate the ability of the IR models to estimate abstract semantic association.

### 4.2 Dataset Construction

**Step 1: Data Collection.** We collect data for training set and test set separately. Specifically, for test set, we use 33 publicly available English books that are collected from Gutenberg project and processed by (Yu et al., 2023). We find 84 Chinese versions of these 33 English books that we have licenses of usage. We sample 52,924 public comments written by readers for various plots in these 84 books. For the training set, we collect 105 books from the same reading app and sample 1,005,480 comments. There is no overlap between books in the training set and the test set.

**Step 2: Data Filtering.** Before human annotation, we perform a preliminary filter on the collected data. Specifically, first, in order to make the description of the comment for the plot abstract enough, we remove the comments that have a lot of word overlap with the original texts in the book. Given a comment $c$ and the original text $t$ in the book marked by the comment $c$, we use NLTK[2] to perform word tokenization on them and remove the stop words. Then we get the sets of words for them ($\mathbb{C}$ and $\mathbb{T}$). We remove the comments that:

$$\frac{|\mathbb{C} \cap \mathbb{T}|}{|\mathbb{C}|} > 0.5. \tag{5}$$

Second, we remove the comments that have little practical meaning. That is, the comments that do not describe the plot but express the reader's emotions such as "*This is so funny!*" or "*I can't understand this*". We use ChatGPT[3] via prompting it to judge whether the comment is describing a specific plot rather than simply expressing emotion to complete this task. Considering that a large amount of data will bring high ChatGPT usage cost, we perform this filtering operation on the full test set and 50,000 samples of training set. For the other samples in the training set, we use the automatic annotation model for fast and low-cost filtering, which will be introduced in Step 5. After this, we get 7,661 samples in test set and 7,432 samples in training set for human annotation.

**Step 3: Human Annotation.** For the sample with a comment $c$ and the original text $t$ in the book marked by the comment $c$, annotators have two tasks to finish. (1) Judge whether $c$ contains the abstract description of the plot in $t$. (2) If so, mark the texts describing the plot from $c$ and use the texts as the query $q$. After this step, we can get the query-plot pairs where there is the abstract semantic association between query and plot. Specifically, we first select nine annotators who have at least a high school education level, because our task requires the annotators to have a certain ability to understand literary works. We write the guidelines to help the annotators better understand the details of the annotation task. Before the formal annotation start, we conduct three rounds of pre-annotation and verify the pass rate of each annotator's work. We select the annotator whose pass rate of work reaches 90% in the pre-annotation for formal annotation. In the formal annotation, for the results of each annotator, we introduce another annotator to sample and validate the results and give the pass rate, which can measure whether two annotators agree with the results. We continue to screen and guide the annotators until the pass rate of each annotator reaches 95%. We select the samples that $c$ are judged to contain abstract descriptions of $t$ as the final samples. After this, we get 4,572 query-plot samples in the test set and 4,402 samples in the training set.

| | |
|---|---|
| #Train Pairs | 400,000 |
| #Validation Pairs | 37,609 |
| #Test Quries | 4,572 |
| #Candidate plot chunks | 136,195 |
| Average query length | 29.12 |
| Average chunk length | 58.10 |

Table 1: Statistics of PLOTRETRIEVAL .

**Step 4: Translation and Corpus Construction.** Since the majority of our collected data is in Chinese, we translate the collected data into English. For test set, all books have their public English versions (Step 1). So we (1) translate the comment $c$ to English and (2) project the original text $t$ in the Chinese book marked by the comment $c$ to its content in the English version of the book. For the first task, we finish it by ChatGPT. For the second task, we use Spacy to sentencize the texts of books, use multilingual embedding LASER[4] to embed sentences and use *vecalign* (Thompson and Koehn, 2019) to align the sentences between books based on sentence embeddings. For training set, because some books do not have the corresponding English versions, we directly translate $c$ and $t$ to English by *Helsinki*[5], a neural machine translation model.

We use the collection of plots of books in the test set as the retrieval corpus, which means that when we test the retrieval performance of the IR models on PLOTRETRIEVAL, the samples in the training set do not appear in any test data. For the book, we divide every $m$ sentences into a chunk (the basic unit of the corpus). We mark the chunks containing the sentences in $t$ as ground truth for $c$. To ensure the semantic integrity of $t$, we also make $t$ as a chunk and mark it as ground truth. Details of the corpus are shown in Appendix B.2.

**Step 5: Auto Annotation Model.** For the large amount of data in the training set that has not been

[2] https://www.nltk.org/
[3] https://openai.com/blog/chatgpt
[4] https://github.com/facebookresearch/LASER.
[5] https://huggingface.co/Helsinki-NLP

| Dataset | Word Overlap |
|---|---|
| FEVER | 61.57 |
| Quora | 53.75 |
| Touché-2020 | 51.77 |
| SCIFACT | 48.24 |
| MS MARCO | 46.29 |
| Dbpedia | 41.54 |
| FiQA-2018 | 38.40 |
| NQ | 36.24 |
| HotPotQA | 35.66 |
| Climate-Fever | 29.02 |
| Arguana | 28.98 |
| SCIDOCS | 26.79 |
| Trec Covid | 26.41 |
| NFCorpus | 23.33 |
| PLOTRETRIEVAL | **19.62** |

Table 2: Word overlap between query and the positive candidate documents among various IR datasets.

manually annotated, we construct a text-pair binary classifier to complete automatic annotation. Specifically, we train BERT[6] (Devlin et al., 2019) on 50,000 samples of training set in Step 2 in which 4,402 are annotated as positives in Step 3 and the other are negatives. We use the trained classifier to automatically annotate the data in the training set. Although most of the data in the training set is constructed under the weak supervision of the automatic annotation model, experiments in Section 5.3 show that compared with large-scale supervised IR datasets, our training data is better for IR models to estimate the abstract semantic association.

### 4.3 Data Statistics

Table 1 shows the statistics of the training set and test set in PLOTRETRIEVAL. Most of the train and validation pairs are obtained from the auto annotation model in Step 5. Table 2 shows the word overlap between the query and candidate documents (calculated by Equ (5)). PLOTRETRIEVAL has the lowest overlap, especially compared to mainstream IR datasets such as MS MARCO. Therefore, compared to the existing IR datasets, the query-plot pairs in PLOTRETRIEVAL pose a higher challenge to the IR models. The pairs look very different but have abstract semantic association, rather than simple lexical or semantic matching.

## 5 Experiments

In this section, we evaluate various IR models on PLOTRETRIEVAL and perform human studies.

[6]https://huggingface.co/bert-base-uncased

### 5.1 Baselines

**Lexical Retrieval.** We use (1) **BM25** (Robertson et al., 1995), a a bag-of-words retrieval method based on word-to-word exact matching.

**Sparse Retrieval.** Following BEIR (Thakur et al., 2021), we select three mainstream sparse retrieval models including (1) **DeepCT** (learning dynamic term weights) (Dai and Callan, 2020), (2) **SPARTA** (learning a sparse representation that can be efficiently implemented as an inverted index) (Zhao et al., 2021) and (3) **DocT5query** (generating queries added to documents) (Nogueira and Lin, 2019). All of them are fine-tuned on MS MARCO.

**Dense Retrieval.** (1) **DPR** (Karpukhin et al., 2020), a classical dense retrieval model based on bi-encoder and trained with BM25 hard negatives and in-batch contrastive loss. (2) **ANCE** (Xiong et al., 2021), it dynamically updates negatives during training. (3) **TAS-B** (Hofstätter et al., 2021) is trained with supervision from cross-encoder. (4) **BERM** (Xu et al., 2023a,b), a plug-and-play method to enable dense retrieval models to learn representations that are more suitable for matching. (5) **Ernie-Search** (Lu et al., 2022) trains dense retrieval model by cascade distillation from ColBERT (Khattab and Zaharia, 2020) and cross-encoder. All of the above baselines are fine-tuned on MS MARCO. There are also some methods first pre-train models on large-scale datasets by self-supervised IR signal. (6) **COCO-DR** (Yu et al., 2022) is pre-trained on BEIR (Thakur et al., 2021). (7) **coCondenser** (Gao and Callan, 2022) and (8) **RetroMAE** (Xiao et al., 2022) are pre-trained on English Wikipedia and BookCorpus. (8) **Contriever** (Izacard et al., 2022) is pre-trained on English Wikipedia and CCNet. All of these models are fine-tuned on MS MARCO after pre-training for IR.

**Late-Interaction.** **ColBERT** (Khattab and Zaharia, 2020) performs late interaction on embeddings of each token to achieve finer-grained interaction than dense retrieval. This model is fine-tuned on MS MARCO.

**Re-Ranking.** We use **Cross-Encoder** (Wang et al., 2020) that exploits self-attention for interaction between tokens as re-ranker, which has shown power in Book QA tasks (Mou et al., 2021). Before re-ranking, we first use Contriever to retrieve Top-100 documents for each query as its candidate list. This model is fine-tuned on MS MARCO.

**ChatGPT-Assisted.** ChatGPT performs well on various NLP tasks, we also explore its performance on *Plot Retrieval*. It is expensive to directly let ChatGPT inference on a large-scale corpus, so we prompt ChatGPT to generate the plot in the corresponding book for the query (query expansion (Carpineto and Romano, 2012)), and then use the generated plot as query and use Contriever to retrieve related plots from the corpus.

### 5.2 Experimental Settings

**First**, to explore the ability of the SOTA IR models trained on MS MARCO to estimate abstract semantic associations between texts, we evaluate the performance of them in zero-shot setting on the English version of PLOTRETRIEVAL. **Second**, to show the effectiveness of our weakly supervised training data, we compare the performance of IR models trained on weakly supervised training data in PLOTRETRIEVAL with existing IR datasets in the same training method and settings. We use *bert-base-uncased* and *bert-base-chinese* as pre-trained models for English and Chinese respectively. In training, we set the learning rate to $10^{-5}$. We train the model with 64 batch size on a single A100 GPU for 5 epochs and use Pytorch (Paszke et al., 2019) as the training framework. **Third**, the difficulty of PLOTRETRIEVAL for IR models can be reflected by the performance gap between IR models and humans on different datasets. We compare this gap on different IR datasets via human studies.

### 5.3 Experimental Results

**Performance on PLOTRETRIEVAL.** Table 3 shows the zero-shot performance of IR models trained on MS MARCO on test set of PLOTRETRIEVAL. We can draw the following four conclusions. **(1)** PLOTRETRIEVAL has more abstract semantic association and less word overlap between texts than existing IR datasets, which is more challenging for current SOTA IR models. This can be supported by the phenomenon that BM25, the strong zero-shot IR baseline based on term-matching (Thakur et al., 2021; Izacard et al., 2022), achieves better performance on BEIR (Thakur et al., 2021) than many neural IR models such as DPR, ANCE, and TAS-B, but has worse performance on PLOTRETRIEVAL than all neural IR baselines that can capture the semantic matching information. **(2)** More training data facilitates the estimation of abstract semantic association, even if the data is self-supervised. This can be supported by the phenomenon that models pre-trained on large-scale datasets such as coCondenser, Contriever, COCO-DR, and RetroMAE have better performance than the models fine-tuned directly on MS MARCO. **(3)** More interactions between texts are conducive to the estimation of abstract semantic association. Cross-Encoder that exploits self-attention for fine-grained interaction between tokens shows the best performance. **(4)** ChatGPT is not good at associating plots with their abstract corresponding queries. Using ChatGPT to generate the plot associated with the query, and using the generated content as the new query for retrieval by Contriever achieves worse performance. It is because we find that ChatGPT cannot accurately generate the plots associated with the query but generates the common content for the book such as the summary and background of the book. This makes the query ambiguous and indiscriminate.

**Discussion on N-RODCG.** In this paper, we propose a new evaluation metric named N-RODCG, which is more in line with the actual book reading scene. Specifically, traditional IR metrics such as MRR, Recall and NDCG can only reflect the difference in relevance between the texts in the returned rankted list and the ground-truth. However, in the book reading scene, a more reasonable metric is to reflect the distance between the retrieved texts and the ground-truth in the book. Because this can better reflect the retrieval models' ability to help readers find the content they want from the book. The greater the value of the metric, the closer the retrieved texts is to the ground-truth in the book, and the easier for readers to find what they want to read.

**Effect of Weakly Supervised Training Data.** The weakly supervised training data we construct has positive significance for improving the performance of the IR models on the task *Plot Retrieval*. Specifically, we compare the performance of models trained on mainstream supervised datasets (human annotation) with the models trained on weakly supervised training data in PLOTRETRIEVAL. In English setting, we use two datasets as baselines. The one is MS MARCO, the large-scale labeled IR dataset. The other is RELiC (Thai et al., 2022), the large-scale labeled IR dataset that aims to retrieve evidence for literary claims, whose domain also involves book reading. In Chinese setting, we use DuReader (Qiu et al., 2022), a large-scale Chinese labeled IR dataset. These models are fine-tuned with the same method (DPR) and settings

| Model | MRR | | | Recall | | | N-RODCG | | |
|---|---|---|---|---|---|---|---|---|---|
| | @1 | @10 | @100 | @1 | @10 | @100 | @1 | @10 | @100 |
| | | | | Lexical Retrieval | | | | | |
| BM25 | 0.063 | 0.093 | 0.100 | 0.063 | 0.083 | 0.182 | 0.077 | 0.085 | 0.125 |
| | | | | Sparse Retrieval | | | | | |
| SPARTA | 0.059 | 0.090 | 0.098 | 0.059 | 0.096 | 0.253 | 0.069 | 0.088 | 0.143 |
| DeepCT | 0.043 | 0.085 | 0.091 | 0.043 | 0.089 | 0.242 | 0.058 | 0.082 | 0.136 |
| docT5query | 0.085 | 0.124 | 0.136 | 0.085 | 0.130 | 0.330 | 0.107 | 0.129 | 0.199 |
| | | | | Dense Retrieval | | | | | |
| DPR | 0.081 | 0.123 | 0.132 | 0.081 | 0.129 | 0.321 | 0.098 | 0.121 | 0.193 |
| ANCE | 0.088 | 0.129 | 0.139 | 0.088 | 0.136 | 0.332 | 0.110 | 0.132 | 0.204 |
| TAS-B• | 0.091 | 0.140 | 0.150 | 0.091 | 0.161 | 0.373 | 0.112 | 0.148 | 0.227 |
| BERM | 0.088 | 0.132 | 0.141 | 0.088 | 0.149 | 0.354 | 0.107 | 0.137 | 0.214 |
| coCondenser⋆ | 0.097 | 0.146 | 0.155 | 0.097 | 0.162 | 0.368 | 0.116 | 0.151 | 0.227 |
| Ernie-Search• | 0.102 | 0.151 | 0.161 | 0.102 | 0.167 | 0.381 | 0.124 | 0.158 | 0.238 |
| Contriever⋆ | 0.111 | 0.165 | 0.175 | 0.111 | 0.184 | **0.416** | 0.137 | 0.176 | 0.262 |
| COCO-DR⋆ | 0.096 | 0.145 | 0.155 | 0.096 | 0.158 | 0.375 | 0.118 | 0.150 | 0.231 |
| RetroMAE•⋆ | 0.108 | 0.158 | 0.168 | 0.108 | 0.174 | 0.395 | 0.132 | 0.168 | 0.249 |
| | | | | Late-Interaction | | | | | |
| ColBERTv2 | 0.120 | 0.170 | 0.179 | 0.120 | 0.144 | 0.290 | 0.141 | 0.151 | 0.211 |
| | | | | Re-Ranking | | | | | |
| Cross-Encoder | **0.123** | **0.174** | **0.184** | **0.123** | **0.197** | **0.416** | **0.150** | **0.189** | **0.272** |
| | | | | ChatGPT-Assisted | | | | | |
| ChatGPT+Contriever | 0.048 | 0.077 | 0.085 | 0.048 | 0.088 | 0.254 | 0.062 | 0.083 | 0.142 |

Table 3: Zero-shot performance of IR models on test set of PLOTRETRIEVAL. **Bold**: best performance. Underlined: second best performance. ⋆: Train on large self-supervised data. •: Knowledge distillation from cross-encoder.

and perform early stopping on validation pairs. Table 4 shows that weakly supervised training data in PLOTRETRIEVAL significantly improves the performance of the IR models on *Plot Retrieval* than mainstream supervised IR datasets with much more human annotations. We maintain the independence of the training set and test set in the process of data construction so that there is enough domain gap between them. Besides, although RELiC also belongs to the book domain, its performance is not significantly improved compared with MS MARCO. This further shows the effectiveness of our weakly supervised training data for IR models to learn the abstract semantic association between texts instead of just overfitting the domain.

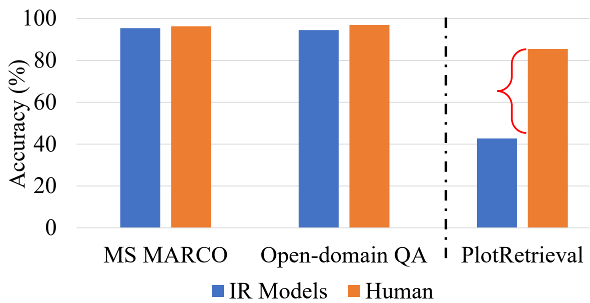


Figure 3: The gap between IR models and humans.

**Human Studies.** We perform human studies to compare the performance gap of IR models and humans on MS MARCO, ODQA (consisting of Natural Questions, TriviaQA, SQuAD, WebQuestions), and PLOTRETRIEVAL. Specifically, we sample 500 queries from the test sets of these three datasets respectively, for each query, we construct a candidate list containing 1 ground truth and 19 negatives. We let the IR model and humans select the ground truth for the query from its candidate list and count the accuracy. We use DPR (trained on MS MARCO) for MS MARCO, DPR (trained on ODQA) for ODQA, and Cross-Encoder (the best model in Table 3 and trained on MS MARCO and PLOTRETRIEVAL ) for PLOTRETRIEVAL as the IR models. We select three humans with college degrees for this study and count the average accuracy. Results in Figure 3 show that although the performance of the IR models on MS MARCO and ODQA is close to human, they still struggle in capturing abstract semantic association on PLOTRETRIEVAL.

| Dataset | Domain | MRR | | | Recall | | | N-RODCG | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | @1 | @10 | @100 | @1 | @10 | @100 | @1 | @10 | @100 |
| | | | | | English Setting | | | | | |
| MS MARCO | Misc. | 0.080 | 0.121 | 0.131 | 0.080 | 0.125 | 0.320 | 0.095 | 0.119 | 0.190 |
| RELiC | Book | 0.083 | 0.128 | 0.138 | 0.083 | 0.142 | 0.389 | 0.102 | 0.134 | 0.225 |
| PLOTRETRIEVAL (weakly supervised) | Book | **0.105**† | **0.155**† | **0.165**† | **0.105**† | **0.174**† | **0.420**† | **0.128**† | **0.163**† | **0.253**† |
| | | | | | Chinese Setting | | | | | |
| DuReader | Misc. | 0.031 | 0.041 | 0.045 | 0.031 | 0.062 | 0.175 | 0.041 | 0.075 | 0.139 |
| PLOTRETRIEVAL (weakly supervised) | Book | **0.103**† | **0.152**† | **0.164**† | **0.103**† | **0.247**† | **0.588**† | **0.140**† | **0.169**† | **0.257**† |

Table 4: Performance of the (DPR) models trained on different IR datasets on test set of PLOTRETRIEVAL. **Bold**: best performance. †: significant performance improvement with p-value $\leq 0.05$ compared with baselines.

## 6 Conclusion

In this paper, we propose a novel task called *Plot Retrieval* that retrieves relevant plots from the book for a query. Compared with the existing IR datasets, *Plot Retrieval* requires the IR models to have the strong ability to capture the abstract semantic association between texts rather than the simple lexical and semantic matching. It is meanly because readers integrate their own understanding, summaries, or speculations of the plot when writing the query. For the *Plot Retrieval* task, we propose PLOTRETRIEVAL, a large labeled dataset with more abstract semantic association and less word overlap between texts, which can be used as a benchmark to train and evaluate the ability of IR models to capture abstract semantic associations between texts. Extensive experiments across various lexical retrieval, sparse retrieval, dense retrieval, and cross-encoder methods compared with human studies on PLOTRETRIEVAL show that the current IR models still struggle in capturing abstract semantic association between texts and there is a lot of room for improvement in future research.

## Limitations

In this paper, we propose a novel task called *Plot Retrieval*. *Plot Retrieval* aims to retrieve the relevant plots for the query and has higher requirement for the ability of the information retrieval models to estimate the abstract semantic association between texts while existing information retrieval datasets are not satisfied. To achieve it, we collect and release PLOTRETRIEVAL, a large-scale information retrieval dataset with more abstract semantic association and less word overlap. However, although comparison with humans shows that current SOTA IR models cannot perform well at this task, we do not propose an efficient solution such as novel model architecture and training method to solve this problem. Our contributions focus on proposing a more challenging retrieval task and dataset. Further research on the task will be carried out in future work.

## Ethics Statement

In the construction of datasets, we prioritize the ethical use of data and are committed to upholding the highest standards when it comes to protecting user privacy and ensuring data integrity. Specifically, all the data within our dataset is collected exclusively from publicly available information from online applications (apps). We strictly adhere to the legal guidelines and terms of service of these apps during the data collection process. Our data collection practices prioritize user privacy. All personally identifiable information (PII) has been thoroughly masked or removed from the dataset. We declare that our work complies with the ACL Ethics Policy.


## Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB3103700, 2022YFB3103704), the National Natural Science Foundation of China (NSFC) under Grants No. 62276248 and U21B2046, and the Youth Innovation Promotion Association CAS under Grants No. 2023111.


## References


Petr Baudis and Jan Sedivý. 2015. Modeling of the question answering task in the yodaqa system. In *Proceedings of the Conference on CLEF 215*, volume

9283 of *Lecture Notes in Computer Science*, pages 222–228. Springer.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the Conference on EMNLP 2013*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1).

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 2020 Conference on ACL*, pages 2270–2282, Online. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 2020 Conference on SIGIR*, pages 1533–1536. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference on NAACL*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. *COLING*.

Lea Fremann. 2019. Extractive narrativeqa with heuristic pre-training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 172–182.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. DBpedia-Entity v2: A test collection for entity search. In *Proceedings of the 2017 Conference on SIGIR*, SIGIR '17, page 1265–1268, New York, NY, USA. Association for Computing Machinery.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 2021 Conference on SIGIR*, pages 113–122. ACM.

Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. 2015. CQADupStack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, ADCS '15, New York, NY, USA. Association for Computing Machinery.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on EMNLP*, pages 6769–6781. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yuxiang Lu, Yiding Liu, Jiaxiang Liu, Yunsheng Shi, Zhengjie Huang, Shikun Feng Yu Sun, Hao Tian, Hua Wu, Shuaiqiang Wang, Dawei Yin, and Haifeng Wang. 2022. Ernie-search: Bridging cross-encoder with dual-encoder via self on-the-fly distillation for dense passage retrieval.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Trans. Assoc. Comput. Linguistics*, 9:1032–1046.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rodrigo Nogueira and Jimmy Lin. 2019. Document expansion by query prediction. In *arXiv preprint*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 2019 Conference on NeurIPS*, pages 8024–8035.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, QiaoQiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5326–5338, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1995. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manag.*, 31(3):345–360.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022. A survey of machine narrative reading comprehension assessments. In *Survey Track of 31st International Joint Conference on Artificial Intelligence (IJCAI Survey Track)*.

Ian Soboroff, Shudong Huang, and Donna Harman. 2019. Trec 2019 news track overview. In *TREC*.

Axel Suarez, Dyaa Albakour, David Corney, Miguel Martinez, and José Esquivel. 2018. A data collection for evaluating the retrieval of related tweets to news articles. In *Advances in Information Retrieval*, pages 780–786, Cham. Springer International Publishing.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. RELiC: Retrieving evidence for literary claims. In *Proceedings of the 2022 Confernence on ACL*, pages 7500–7518, Dublin, Ireland. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 2021 Conference on NeurIPS Datasets and Benchmarks*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, and Dimitris Polychronopoulos. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition.

Ellen M. Voorhees. 2004. Overview of the TREC 2004 robust track. In *Proceedings of the Conference on TREC 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 2018 Conference on ACL*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Mengting Wan, Rishabh Misra, Ndapandula Nakashole, and Julian McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2605–2610.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the 2021 Conference on ICLR*. OpenReview.net.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2022. Match-prompt: Improving multi-task generalization ability for neural text matching via prompt learning. In *Proceedings of the Conference on CIKM 2022*, CIKM '22, page 2290–2300, New York, NY, USA. Association for Computing Machinery.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023a. Berm: Training the balanced and extractable representation for matching to improve generalization ability of dense retrieval.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023b. BERM: Training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6620–6635, Toronto, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018a. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on EMNLP*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zhou Yang, Qingfeng Lan, Jiafeng Guo, Yixing Fan, Xiaofei Zhu, Yanyan Lan, Yue Wang, and Xueqi Cheng. 2018b. A deep top-k relevance matching model for ad-hoc retrieval. In *Proceedings of the 2018 Conference on CCIR*, volume 11168 of *Lecture Notes in Computer Science*, pages 16–27. Springer.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *CoRR*, abs/2210.15212.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

# A Interface for Annotation

Figure 4 shows the interface for annotation. The original interface is in Chinese, we translate it into English for better reading.

# B Details of PLOTRETRIEVAL

## B.1 Examples in PLOTRETRIEVAL

Figure 5 shows some examples in PLOTRETRIEVAL.

## B.2 Books in PLOTRETRIEVAL

Table 5 shows the book name in the corpus of test set and the number of queries and plot chunks for each book.

# C Case Study

Table 6 shows the comparison of ground truth with Top-1 results retrieved by Contriever and BM25 respectively. The results of BM25 show that BM25 are limited to word overlap but cannot capture semantic level information. For the results of Contriever, they are limited to literal semantic matching, Contriever cannot deeply understand the meaning that the query really wants to express to find the most suitable plot.

## Interface for Chinese Annotators

### 故事原文:

没有一个不愿为佩格蒂先生效劳的，而且也没有一个被请帮忙的人不得到好好酬谢的，所以帮忙的人有的是，可是葛米治太太整天执意要搬运那些重得她力不胜任的东西，还不辞辛苦地跑来跑去忙着干那些不需要她去干的差使。甚至悲叹她自己的不幸，她好像也完全忘了，不记得自己有过任何不幸了。她在同情中自始至终保持着乐观的态度，这也是她所起的变化中令人吃惊的一部分。怨天尤人的情况绝对没有了。在那一整天里，我甚至没有听到过她声音打颤，也没有看到过她流过半滴眼泪。

### 评论:

葛米治太太在佩格蒂先生遭遇不幸后，反而不同往常的调整自己的情绪，努力使这个家运转起来，表现出她是一个善解人意的坚强女人，也是一个懂得感恩的女人。

### 任务一：评论是否包含对剧情的描述?

● 包含剧情
○ 不包含剧情

### 任务一：若评论描述了故事原文的剧情，请将评论中描述剧情的文本划出

葛米治太太在佩格蒂先生遭遇不幸后，反而不同往常的调整自己的情绪，努力使这个家运转起来，表现出她是一个善解人意的坚强女人，也是一个懂得感恩的女人。

## English Translation

### Original texts in Book:

There was no one who would not be of service to Mr. Peggotty, and no one who was asked to help was not well paid, so there were plenty of help, but Mrs. Gummidge insisted all day on moving things that were too heavy for her to handle. stuff, running around and doing errands that didn't require her to do. Even bemoaning her own misfortunes, she seemed to have completely forgotten, and could not remember any misfortunes of her own. Her sympathetic optimism throughout is part of the astonishing change she has made. Absolutely no more complaints. During that whole day, I didn't even hear her tremble, and I didn't even see her shed a single tear.

### Comment:

After Mrs. Peggotty's misfortune, Mrs. Gemidge adjusted her emotions differently and tried her best to make the family work, showing that she is a strong woman who understands others and is also a woman who knows how to be grateful.

### Task 1: Does the comment contain description of the plot in the original texts?

● Yes
○ No

### Task 2: If the comment describes the plot of the original texts, please mark the texts describing the plot in the comment.

After Mrs. Peggotty's misfortune, Mrs. Gemidge adjusted her emotions differently and tried her best to make the family work, showing that she is a strong woman who understands others and is also a woman who knows how to be grateful.

Figure 4: Interface for annotation.

| Query | Plot |
|---|---|
| The original intent was good, but then it evolved differently and distorted. | No fight could have been half so terrible as this dance. It was so emphatically a fallen sport--a something, once innocent, delivered over to all devilry--a healthy pastime changed into a means of angering the blood, bewildering the senses, and steeling the heart. Such grace as was visible in it, made it the uglier, showing how warped and perverted all things good by nature were become. |
| The peasants are still too ignorant to understand Nekhludoff's plan. | "So you have enough land?" asked Nekhludoff.<br>"No." The old soldier pretended to be happy and said. He clutched his battered hat to his chest with all his might, as if offering it to someone who would wear it.<br>"However, you must think carefully about what I have said," said Nekhludoff, who was astonished, repeating his suggestion.<br>"We don't have to think about it. We do what we say," said the toothless, sullen old man angrily. |
| These people are so indifferent to pain, death and suffering. | They were not allowed near the carriages. Escorts are particularly worried today. Along the way from the prison to the station, in addition to the two Nekhludoff saw, three more prisoners died of heatstroke. One of them, like the first two, was sent to a nearby police station. Fallen at the station. What the escorts were worried about was not that five prisoners who could have been saved died under their escort. They don't take it to heart at all. All they worry about is the dead, and they have to go through various formalities according to the law: send the dead, their materials, and clothing to the relevant departments, and check off their names from the list of prisoners escorted to the lower city. |
| Treat marriage as a kind of atonement, as a kind of self-sacrifice. | "First," he thought, "I'll go to the lawyer now and ask him about his decision, and then...then I'll go to the prison to visit yesterday's female prisoner and tell her everything."<br>He imagined that he would visit her, tell her everything before and after, admit his fault to her, and tell her solemnly that he would do his best for her, and that he would marry her to atone for his sin. Thinking of this, his heart was filled with special joy, and tears welled up in his eyes. |
| Female prisoner attracts the attention of people on the street. | Two soldiers escorted the female prisoner down the steps toward the gate. A small door above the gate was opened, and two soldiers escorted the prisoner across the threshold into the courtyard, then out of the courtyard wall into the stone-paved street in the middle of the city. The coachman, shop owner, cook, worker, and official all stopped and looked at the female prisoner curiously. Some shook their heads, thinking to themselves: "Look, this guy is not behaving like us, and it's what he did." The children looked at the female prisoner in horror. |

Figure 5: Examplse in PLOTRETRIEVAL .

| Book Name | #Queries | #Plot Chunks |
|---|---|---|
| *The Red and the Black* | 666 | 4353 |
| *The Count of Monte Cristo* | 200 | 9013 |
| *The Adventures of Tom Sawyer Complete* | 121 | 1759 |
| *David Copperfield* | 153 | 6552 |
| *The Gadfly* | 134 | 2426 |
| *A Tale of Two Cities* | 325 | 2911 |
| *Crime and Punishment* | 404 | 5187 |
| *The Brothers Karamazov* | 217 | 8251 |
| *Les Miserables* | 317 | 12030 |
| *Eugenie Grandet* | 126 | 1392 |
| *Tess of the d'Urbervilles* | 343 | 3035 |
| *Notre-Dame de Paris* | 510 | 4270 |
| *The Call of the Wild* | 163 | 729 |
| *The Idiot* | 122 | 5480 |
| *Moby Dick; or The Whale* | 125 | 3429 |
| *Resurrection* | 647 | 3901 |

Table 5: Books in the corpus of test set.

| Query | Ground Truth | Contriever | BM25 |
|---|---|---|---|
| Dana realized Kartun's intentions, but it was too late. Kartun used drugs or something similar to forcibly knock him out. | If it had been otherwise--'" Carton looked at the pen and saw it was trailing off into unintelligible signs. Carton's hand moved back to his breast no more. The prisoner sprang up with a reproachful look, but Carton's hand was close and firm at his nostrils, and Carton's left arm caught him round the waist. For a few seconds he faintly struggled with the man who had come to lay down his life for him; but, within a minute or so, he was stretched insensible on the ground. | You apprehend me very clearly, Mr. Barsad. I won't." Carton's negligent recklessness of manner came powerfully in aid of his quickness and skill, in such a business as he had in his secret mind, and with such a man as he had to do with. | Looking gently at him again, she was surprised and saddened to see that there were tears in his eyes. There were tears in his voice too, as he answered: "It is too late for that. I shall never be better than I am. |
| He hurt himself but made Tess swear not to tempt him again. | "Relics are not in my creed; but I fear you at moments--far more than you need fear me at present; and to lessen my fear, put your hand upon that stone hand, and swear that you will never tempt me--by your charms or ways." | At breakfast, and while they were packing the few remaining articles, he showed his weariness from the night's effort so unmistakeably that Tess was on the point of revealing all that had happened; | In a very few minutes after, he was driving up the hill out of the town which, three or four months earlier in the year, Tess had descended with such hopes and ascended with such shattered purposes. Benvill Lane soon stretched before him, its hedges and trees purple with buds; but he was looking at other things, and only recalled himself to the scene sufficiently to enable him to keep the way. |
| Tom found these clues and marks while being trapped in the cave. He has good psychological quality of remaining calm and composed even when in a difficult situation. | He held his candle aloft and said:<br>"Look as far around the corner as you can. Do you see that? There--on the big rock over yonder--done with candle-smoke."<br>"Tom, it's a cross!"<br>"Now where's your Number Two?<br>'under the cross,' hey?<br>Right yonder's where I saw Injun Joe poke up his candle, Huck! | Tom kissed her, with a choking sensation in his throat, and made a show of being confident of finding the searchers or an escape from the cave; then he took the kite-line in his hand and went groping down one of the passages on his hands and knees, distressed with hunger and sick with bodings of coming doom. | When she found the entire fence white-washed, and not only whitewashed but elaborately coated and recoated, and even a streak added to the ground, her astonishment was almost unspeakable. She said: "Well, I never! There's no getting round it, you can work when you're a mind to, Tom. |
| This is a contradictory personality, loving and hating, despising and appreciating for Julien. | She adored him, and nevertheless she exhibited for a good quarter of an hour in her invective against his, Julien's, character, and her regret at having ever loved him, the same haughty soul which had formerly overwhelmed him with such cutting insults in the library of the Hotel de la Mole. | This unique person never thinks for a minute of seeking help or support in others! He despises others, and that is why I do not despise him. "If Julien were noble as well as poor, my love would simply be a vulgar piece of stupidity, a sheer mesalliance; I would have nothing to do with it; it would be absolutely devoid of the characteristic traits of grand passion--the immensity of the difficulty to be overcome and the black uncertainty cf the result." | M. de Renal's face cleared.<br>"It would also be a black mark," continued Julien in a more humble tone, "against a poor theology student if it ever leaked out that his name had been on the ledger of a bookseller who let out books. The Liberals might go so far as to accuse me of having asked for the most infamous books." |

Figure 6: Comparison between ground truth and Top-1 results of Contriever and BM25.