Automatic Quote Attribution in Literary Works

Xingxing Yang
Division of Emerging Interdisciplinary Areas
Hong Kong University of Science and Technology
China
xyangbx@connect.ust.hk

Yu Wang
Department of Computing Science
University of Alberta
Canada
wang.yu@ualberta.ca

Abstract

In fiction, quote attribution pertains to the process of extracting dialogues and identifying the speakers involved. This encompasses quotation and speaker annotation. To accomplish this, we have developed a pipeline for quote attribution that incorporates classification, extractive question answering (QA), multi-choice QA, and coreference resolution. Additionally, we evaluated our model's performance by employing various models to predict both explicit and implicit speakers.

1 Introduction

Quote attribution, within the realm of literature and textual analysis, plays a pivotal role in enhancing the clarity and understanding of dialogues. It involves the extraction of dialogues from a text and the subsequent identification of the speakers involved. By assigning the appropriate speakers to their respective utterances, quote attribution enables readers to follow the flow of conversation and comprehend the nuances of a narrative. While manual quote attribution has been the traditional approach, advancements in natural language processing and machine learning have opened up exciting possibilities for automating this process. In this article, we explore the challenges and techniques associated with quote attribution, as well as the significance of automated methods in facilitating efficient and accurate analysis of dialogues in various literary contexts.

We leverage an annotated dataset consisting of 1991 modern Chinese novels as the foundation of our research. While our primary focus is on modern Chinese literature, our methodology can be seamlessly applied to other languages as well, despite potential language differences. Initially, our approach revolves around treating quote attribution as an extractive question-answering (QA) problem. To accomplish this, we fine-tune a pre-trained BERT model specifically for dialogues within fictional texts.

As we delve deeper into the complexities of quote attribution, we address more intricate scenarios, such as anaphora and the continuity of conversational threads(Muzny et al., 2017). To effectively handle the resolution of conversational threads, we employ a pre-trained BERT model designed for multi-choice QA. Furthermore, we utilize co-reference resolution techniques to tackle anaphoric references within the dialogues.

In order to distinguish between crowds, soliloquies, and dialogues, we employ a combination of rule-based filtering and a BERT-based classifier. This hybrid approach enables us to accurately identify and categorize different speech patterns and formats within the text.

Our research encompasses several contributions, which can be summarized as follows:

1. We conduct a thorough analysis of the intricate complexities and nuances associated with accurately attributing quotes.

2. We introduce a comprehensive pipeline for quote attribution, comprising multiple stages including classification, extractive question answering (QA), multi-choice QA, and coreference resolution. This pipeline serves as a systematic framework for effectively and efficiently handling quote attribution tasks.

2 Background/Related Work

To perform quote attribution, a dataset containing quoted speech from literary texts is essential for training, evaluation, and testing models. Several studies have focused primarily on creating datasets specifically for quote attribution. There are some open-source datasets available in this field, most of which are English literature works. Muzny, et al. developed an annotation tool for quotation annotation and created the QuoteLi3 dataset with speaker

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 1–9
August 16, 2024 ©2024 Association for Computational Linguistics

and quotation annotations in 3 novels (Muzny et al., 2017). Similarly, Sims, Matthew, et al. presented datasets for speaker attribution, comprised of 1,765 quotations linked to their speakers in 100 different literary texts (Sims et al., 2019), which is now part of the LitBank corpus. The PNDC project has undertaken similar work (Vishnubhotla et al., 2022). Notably, in these datasets, coreference information is also annotated.

As a crucial first step for quote attribution, quotes need to be extracted from the main text. According to the findings by O'Keefe et al., utilizing regular expressions for quote detection can achieve an accuracy exceeding 99% on clean English language data (O'Keefe et al., 2012).

Concerning quote attribution, multiple approaches have been explored in the existing literature. Earlier works typically employed rule-based methods with grammatical rules, complemented by some machine learning techniques, which could be complex due to the need for defining extensive rules (Elson and McKeown, 2010; Krug, 2020). In Pan et al.'s novel understanding system in 2021, they treated speaker identification of dialogues (SID) as a GBDT-based ranking task. It involved first identifying characters using NER, then feeding input features of the candidate speaker and the target dialogue, such as position, distance, etc., into a model for classification to determine the final speaker from the candidates (Pan et al., 2021). To resolve coreferences, they split names into name clusters with primary and candidate names, extracted features like gender, overlapping, personal pronouns, etc., and used a GBDT-based model with name candidates for coreference resolution. LG Pang employed BERT and CRF (Conditional Random Field), framing the task as question answering, for speaker extraction of quotations[1]. Yoder, Michael Miller, et al. and Vishnubhotla, Krishnapriya, et al. approached quotation attribution as a set of sequential sub-tasks: character identification, coreference resolution, quotation identification, and speaker attribution (Vishnubhotla et al., 2023; Yoder et al., 2021).

In contrast to the recent work by Vishnubhotla, Krishnapriya, et al., we propose an additional module for classifying crowds, soliloquies, and dialogues. We omit the character identification module. Instead, we employ a question-answering (QA) approach for quote attribution, which directly predicts the name of the character speaking the quote.

3 Dataset

We utilize a dataset that we refer to as CLD, comprising 1,991 modern Chinese novels annotated by literature practitioners from the audiobook industry, ensuring high accuracy and validation. The original dataset includes information about the main characters and annotated novel texts, with each quote attributed to a specific character. For every character, we have access to their gender information and a brief description.

4 Quote Attribution

Our primary objective is to correctly attribute each quote to the appropriate speaker within the novel. The process begins with quote extraction, where quotes are identified and extracted from the text. Subsequently, we aim to assign the extracted quotes to their corresponding speakers accurately.

We identify several key challenges in this task:

1. Coreference: Characters in literary texts usually appear in three formats: proper names (e.g., "夏洛克·荷马(Sherlock Holmes)"), pronouns (e.g., "他(He)"), and nominal anaphoric noun phrases referring to characters (e.g., "侦探(The consulting detective)") (Labatut and Bost, 2019). The first-person pronoun "我" (I/My) also frequently appears in some first-person novels.

2. Crowds/System/Sound Effects: There are occasions when the quotes are not spoken by a specific individual but rather by crowds, a "system" (which frequently occurs in certain Chinese time-travel novels), or represent audio effects rather than human speech.

3. Following Conversational Threads: Dialogues often follow an "ABAB" pattern, where A denotes one speaker and B denotes another. Many times, there may be no explicit names or references to the speakers within the paragraph, with only the utterances themselves present.

4. Long Soliloquies: Multiple continuous utterances may be given by a single speaker.

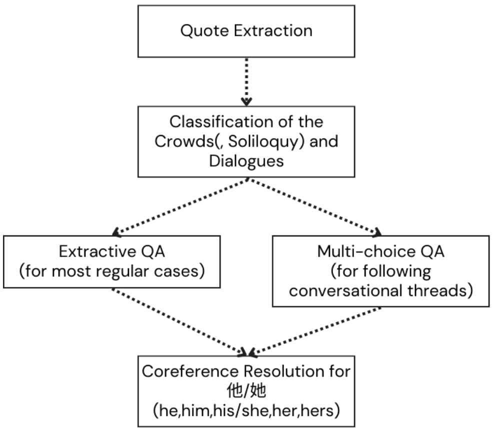
We list examples of different cases in quote attribution in Table 1.

To address the above-mentioned challenges, we employ multiple methods within our proposed pipeline, which can be outlined as follows:

[1] https://gitlab.com/snowhitiger/speakerextraction

Table 1: Different cases in quote attribution

Regular	“好好，还需要什么？”杨建国连忙问。 "Alright, what else do you need?" **Yang Jianguo** asked quickly.
Co-reference	大家都幸灾乐祸地望着叶君临，等待他的回复。 他只得笑笑，断然拒绝：“假酒伤身，我喝不了！” Everyone looked at **Ye Junlin** with schadenfreude and waited for **his** response. **He** could only smile and firmly refuse: "Fake alcohol is harmful to the body, I can’t drink it!"
Sound effects	“咚咚咚……” 就在这时，门外传来一阵清脆的敲门声。 "Dong dong dong..." Just then, **a crisp knocking sound** came from outside the door.
Following conversational threads	“活啥呀，都死九年了，我看咱们是活见鬼！”杨建国哆嗦着说。 “别瞎说，今晚那个玉瑶，不还施法救咱儿子呢吗，谁听说过鬼还捉鬼哩？” 杨妻白了杨建国一眼。 “怪呀，怪呀。不过不管她是啥，咱们都该感谢人家。” “是呀，我还听那小姑娘，管咱家儿子叫小老公……” “要是能再见到这个玉瑶就好了，一定要问问清楚。” 夫妻二人没想到，迷一样的玉瑶，第二天竟然又出现了。 "What kind of life is this? He’s been dead for nine years. I think we’re seeing ghosts!" **Yang Jianguo** shuddered. "Don’t talk nonsense. Didn’t that Yu Yao use magic to save our son tonight? Who’s ever heard of ghosts catching ghosts?" **Yang’s wife** gave **Yang Jianguo** a glare. "It’s strange, it’s strange. But no matter who she is, we should be grateful to her." "Yes, and I heard that little girl calling our son ‘little husband’..." "If we could see this Yu Yao again, it would be great. We have to ask her some questions." The couple didn’t expect that the mysterious Yu Yao would appear again the next day.
Crowds	一掌镇压！而对此，凌风微微抬头，神色风轻云淡，在所有人不可思议的注视下，一根手指缓缓的点出，看上去轻柔而无力。 “这小子在干吗，等死吗？” “我估计是傻了，孟超然师兄可是施展了一品神通，镇山掌。” “狂徒而已，这样轻柔的手指，我上去都可随意灭他，何况孟师兄。” 议论澎湃，只当凌风为小丑。只是，在所有人声音落下的瞬间，前方的一幕，却是让得他们立刻紧闭上了嘴巴，身体剧烈的颤抖起来。 With a single palm strike, Ling Feng suppressed his opponent. However, he lifted his head slightly, his expression calm and relaxed. In the midst of everyone’s incredulous gaze, he slowly pointed a finger, appearing gentle and powerless. "What is the buddy doing, waiting to die?" "I think he’s gone crazy. Senior Meng Chaoran used a first-grade divine technique, the Mountain Suppression Palm." "He’s just a madman. I could easily kill him with a gentle finger like that, let alone Senior Meng." There was **a heated discussion**, and they all thought Ling Feng was a clown. However, in the instant when everyone’s voices fell, the scene in front of them caused them to immediately shut their mouths and their bodies trembled violently.
Long Soliloquy	病床很快推到了病房，这层是高级vip病房区，院长亲自过问，整个楼层都被清空，就安排了颜汐、颜清和还有已经能下地行走的顾念风住。祁愿没有跟进去，而是放手站在了门外，他凝眸看着病房许久，才缓缓道： “祁承，走吧。” “调动人手来守住这里，没有我的同意，连只苍蝇也不准放进来！” “联系国外势力，给他们一天时间，我要颜允之毫发无伤地回到华国！” “放弃攻击颜氏集团，控制舆论消除对颜家的负面影响。傅家霍家那些人谁有意见让他们直接来见我。”…… The hospital bed was quickly pushed into the ward. This floor was the high-level VIP ward area, and the hospital dean personally inquired about the situation. The entire floor was cleared, and Yan Xi, Yan Qing, and Gu Nianfeng, who was already able to walk, were accommodated. **Qi Yuan** did not follow them in, but let go and stood outside the door. **He** stared at the ward for a long time before slowly saying: "Qi Cheng, let’s go." "Arrange manpower to guard this place. Without my permission, not even a fly is allowed in!" "Contact foreign forces and give them one day to ensure that Yan Yunzhi returns to Hualand unharmed!" "Give up attacking the Yan family’s group, control public opinion, and eliminate the negative impact on the Yan family. If anyone from the Fu family or the Huo family has any opinions, let them come see me directly."...


Quote Extraction
Classification of the Crowds(, Soliloquy) and Dialogues
Extractive QA (for most regular cases)
Multi-choice QA (for following conversational threads)
Coreference Resolution for 他/她 (he,him,his/she,her,hers)

4.1 Quote Extraction

As an initial approach, we employ regular expressions, as described by (O'Keefe et al., 2012), for quote extraction. It's important to note that not all authors follow the conventional practice of enclosing speech within quotation marks. However, this method proves effective for the majority of novels in our dataset, and we currently exclude edge cases from our research scope. Consequently, we utilize regular expressions to extract quotes enclosed within symbols such as "", “”, and ‘’, which are commonly used in Chinese fiction.

In some cases, quotes in Chinese novels may represent sound effects, such as "呼呼" (Huhu), which signifies the sound of wind blowing. These sound effect quotes often occur within the same sentence and are typically followed by the word "声" (sound). To identify and filter out these sound effect quotes, we have developed a specific rule. According to this rule, if a quote is less than 10 characters long and has "声" as a suffix, or if it is less than 10 characters and contains no punctuation marks within the quote itself or before and after the quotation marks, it is considered a sound effect quote. However, it's important to note that while this rule successfully handles most cases, it may not cover all possible scenarios.

4.2 QA-based Quote Attribution

For the task of quote attribution, we employ an extractive question-answering (QA) approach. The method is relatively simple yet effective. We construct each data entry in our dataset following the format outlined in Table 2.

When constructing our extractive QA dataset for fine-tuning, we first extract context and speaker pairs with labeled names that have appeared in the given context. This allows the extractive QA model to identify and extract the explicit names along with their start and end indices within the context.

For first-person novels, we treat them as a special case during our dataset preprocessing. Specifically, we consider the pronoun "我" (I) as a character name. This approach is taken because "我" (I) typically refers to the same person throughout the entire novel, even if a specific name is assigned to the character. Resolving the coreference of "我" (I) using name-based resolution can be challenging, as the assigned name often appears only in the initial chapters of the novel.

Moreover, we have observed a performance decrease when the context contains pronouns such as "他" (he) or "她" (she). It is common for these pronouns to be prevalent in certain novels, with the actual character name being distant from the quoted text. To mitigate this issue, we randomly replace character names with "他" (he) or "她" (she) during dataset preprocessing. This approach makes it easier for the model to identify and resolve instances of "他" (he) or "她" (she). Consequently, when handling contexts containing numerous occurrences of these pronouns, we first extract them as names and subsequently apply coreference resolution.

After completing all the previous steps, we proceed to fine-tune a pre-trained RoBERTa model using our prepared dataset.

4.3 MC-based Quote Attribution

The model exhibits performance degradation on extended conversational threads following an ABAB pattern. This is attributed to limitations of the underlying base model in handling longer text sequences. To mitigate this issue, we introduce a supplementary multi-choice model. The underlying assumption is that by restricting the response space to a predefined set of options, we can enhance the accuracy of answer selection.

To improve the model's ability to handle conversational threads, we built a specific training dataset. We prioritized quotes lacking context, focusing on those within a single paragraph. We then expanded by including nearby paragraphs until the speaker was identified. This ensures the dataset captures complete conversational exchanges, empowering the new multi-choice BERT model to handle them effectively.

After constructing each data pair in the format shown in Table 3, we fine-tune the multi-choice BERT model for improved performance.

Table 2: QA-based quote attribution

Context:	见许念念一脸呆滞，杨翠花哭得更伤心，一把鼻涕一把眼泪的抱着她，肥胖的身体哭得不停颤抖。 “孩子她爹，我们念念要是不行，我也不活了……我苦命的儿啊……” 许念念是被这最后的尖锐声音给刺激回神的。 Seeing Xu Nian Nian’s dull face, **Yang Cuihua** cried even more sadly, hugging Niannian with snot, with tears streaming down her plump body that were trembling with non-stop sobs. **"Child’s dad, if Nian Nian dies, I won’t be able to live either... my poor child..."** Xu Nian Nian was brought back to reality by the sharp last sentence.
Question:	“孩子她爹，我们念念要是不行，我也不活了……我苦命的儿啊……”是谁说的? Who said "Child’s dad, if Nian Nian dies, I won’t be able to live either... my poor child..."?
Answer:	杨翠花 Yang Cuihua

Table 3: MC-based quote attribution

Context:	说着，杨天明就把骨头扔了出去。 铁柱打了个激灵：“你不说这是龙坟吗？” “猜的……” 铁柱无语：“现在咋整？” “走吧，天也快黑了，最好赶在天黑之前下山，大人们都说潜山上闹鬼的。” 铁柱快哭了：“来之前你可没这么说。” “我也是才想起来的嘛。”杨天明无辜地摊了摊手。 After speaking, **Yang Tianming** threw the bone out. Tie Zhu shuddered: "Don’t you say this is the Dragon Tomb?" "Guess..." Tie Zhu didn’t know what to say: "What’s going on now?" **"Let’s go, it’s getting dark soon Now, it’s best to go down the mountain before dark, the adults said that the hidden mountain was haunted.”** Tie Zhu was about to cry: “You didn’t say that before we came here.” “I just remembered it too. "Yang Tianming spread his hands innocently.
Question:	“走吧，天也快黑了，最好赶在天黑之前下山，大人们都说潜山上闹鬼的。”是谁说的? Who said "Let’s go. It’s getting dark. It’s best to get down the mountain before it gets too dark. The adults all say that there are ghosts on the mountain,"?
Choices:	[杨天明，铁柱] [Yang Tianming, Tiezhu]
Answer:	杨天明 Yang Tianming

4.4 Co-reference Resolution

Novels use many references to connect ideas and characters. In our case, as we’re focused on who said what (quote attribution), we only care about references that identify speakers within quotes.

Co-reference resolution in fiction faces unique challenges. First, references can span long passages, making it difficult to keep track of who is being referred to. Second, pronouns can shift between characters within a limited context, further confusing the interpretation. These factors contribute to the difficulty, even for humans, of identifying the correct referent. Table 4 showcases some common scenarios where co-reference becomes ambiguous.

We construct our gender pronoun resolution dataset through a combination of automation and human validation. We begin by automatically extracting quote paragraphs containing relevant pronouns. We then enrich the context by including surrounding paragraphs until the speaker is identified. Utilizing character information, we assign the closest speaker of the same gender to the pronoun. However, to ensure accuracy, human validators meticulously review and refine the automatically generated labels, guaranteeing a high-quality dataset for our task.

Our goal is to link pronouns within quotes to the speaker’s name, directly in the original text. To achieve this, we fine-tune a co-reference resolution model. By improving this model’s accuracy, we can precisely connect pronouns to speakers, enabling a smoother integration with our existing extractive question answering approach.

We specifically leverage the method in Fast-

Table 4: Different Cases in Co-reference Resolution

Easily recognizable context	但凌天并没有任何恐惧，毕竟，他小时候跟着吴中胤没少来了这个地方。 他不免叫了一声：“青青姐，你在这儿吗？” But $LingTian_1$ didn’t have any fear. After all, he_1 followed Wu Zhongyin to this place a lot of times when he_1 was a child. He_1 couldn’t help calling: "Sister Qingqing, are you here?"
Multiple referees for a single pronoun	刚接过来，上官香放到了桌上，迫不及待地就吃了起来，小桃看着她狼吞虎咽的模样，担心她会噎着，叮嘱她慢点吃。 “这是我应该做的，公主。”念念不忘慕容夫人对她的叮嘱，要照顾好上官香，千万不能让她饿着。 Just as she_1 received it, $ShangguanXiang_1$ placed the pastry on the table and eagerly began to eat. $Xiaotao_2$ watched her_1 wolfing it down and worried that she_1 might choke, so she_2 advised her_1 to eat slowly. "It’s what I should do, Your Highness," she_2 said, remembering Lady Murong’s instructions to take care of $ShangguanXiang_1$ and not let her_1 go hungry.
Unrecognizable context by human	冥心大帝目光深邃，盯着不断轮动的画面，掌心里多出一件奇特的物件，开口道：“差不多了。” “什么？”司无涯生出一种不太好的感觉。 他语气一沉，继续道，“此物名为天道大璋，蕴含天地规则……是勾连十大规则的关键至宝。” 接着…… $EmperorMingxin_1$ gazed deeply at the constantly rotating screen, there was a strange object in his palm, and said: "It’s almost there." "What?" $SiWuya_2$ had a bad feeling. $His_?$ tone sank, and $he_?$ continued, "This thing is called Tiandao Dazhang, and it contains the rules of heaven and earth... It is the key treasure that connects the ten rules." Then……

coref[2](Toshniwal et al., 2020a), which employs a bounded memory approach to prioritize the most crucial parts of a document and disregard irrelevant information. It’s important to distinguish our approach from prior work in other languages; we replace the original pre-trained Longformer model with its Chinese counterpart and incorporate Chinese word segmentation for task compatibility.

4.5 Classifying the Crowds, Soliloquy, and Dialogues

We’ve identified three common scenarios where continuous utterances are likely to occur: crowds, soliloquies (internal monologues), and dialogues with multiple speakers. For the case of crowds, these "group quotes" often appear as a series of continuous utterances containing keywords like "everyone," "several people," or "the whole class".

We first extract such cases using our annotated data. In our annotated data, crowds are usually labeled with "龙套"(others). It usually comes in the format of at least 3 continuous utterances.

We compared the performance of rule-based method and the BERT method for classification of the crowds, soliloquy and dialogues.

4.5.1 Rule-based Method for Classifying the Crowds

To address such cases, we implement a rule-based approach that involves maintaining a predefined list of keywords for filtering out irrelevant content. These keywords are derived through data analysis. In our annotated dataset, utterances attributed to crowds are typically labeled as "龙套" (others). For analysis purposes, we extract data consisting of a minimum of three consecutive utterances, each containing only a quotation without any accompanying context, and all labeled as "龙套" (others). This dataset subset allows us to perform detailed analysis and further refine our keyword list for effective filtering.

Based on our analysis, we compile a list of keywords that includes terms like "议论" (discussion), "众人" (the crowds), and others. During the inference phase, we extract a minimum of three consecutive utterances with the closest context and examine the surrounding context (excluding the utterances themselves) for the presence of any of the keywords. This approach helps us identify passages that are likely attributed to crowd dialogue.

4.5.2 BERT Classification

In addition to the rule-based method, we also incorporate a BERT-based approach for comparison. In

[2] https://github.com/shtoshni/fast-coref

this approach, we consider the identification of continuous utterances as a three-category classification problem: crowds, soliloquies, and dialogues. Similar to the rule-based method, we extract data for these three cases. The construction of the dataset follows a similar process, with the distinction that we do not rely on predefined keywords for filtering purposes. Instead, the BERT-based method leverages the power of the model to learn and classify utterances based on their contextual information.

Furthermore, we conduct a comparison between the results of binary classification, where the focus is on distinguishing between the two-category classification approach and the three-category one.

5 Experimental Setup

5.1 Text Preprocessing

We construct our data for QA-based quote attribution as shown in Table 2, for MC-based quote attribution as shown in Table 3, and for co-reference resolution as shown in Table 4.

5.2 Training

In all experiments, we use the same original dataset which contains 1991 novels. For each separate task, we correspondingly pre-process the dataset.

To perform QA-based quote attribution, we finetune a Roberta-based QA model using the QA pipeline in the UER (Universal Encoder Representations) toolkit[3](Zhao et al., 2019). In this process, we utilize approximately 16,000 records and structure our data according to the format presented in Table 2. To ensure comprehensive evaluation, we conduct separate experiments for both single-paragraph and mixed-paragraph contexts. For the mixed-paragraph context, we take care to ensure diversity by including 60% single-paragraph instances and 40% multi-paragraph instances in our dataset. This approach allows us to assess the model's performance under different contextual scenarios.

To facilitate MC-based quote attribution, we finetune an MC model using the UniMC framework using the fenshen framework[4](Wang et al., 2022). When constructing our dataset, we follow the format outlined in Table 3. This approach allows us to train the model using multiple-choice questions and corresponding answer options, enabling it to effectively attribute quotes.

For co-reference resolution, we use fast-coref[5] (Toshniwal et al., 2021, 2020b) and replace the English base model with the Chinese pre-trained Longformer model[6] and used Jieba[7] for Chinese word segmentation. We use around 7000 records.

To perform classification tasks for crowds, soliloquies, and dialogues, we employ BERT (Bidirectional Encoder Representations from Transformers)[8] (Devlin et al., 2018) as our classification model. For each category, we utilize a dataset consisting of 2000 records to train the model.

6 Results and Discussion

To evaluate the performance of these models, we randomly extract 10% of the dataset for evaluation for each model. Here are our results.

Table 5: Results of extractive QA for quote attribution

Case	F-score	EM
Single-paragraph	95.1794	93.1452
Mixed(Single- & Multi-paragraph)	88.2331	86.3062

Table 6: Results of MC for quote attribution

Model	Accuracy
UniMC	0.9259

Table 7: Results of co-reference resolution

Model	F-score
Fast-coref	95.4

The results indicate that individual modules achieve high performance, showcasing the promising accuracy for quote attribution. We didn't conduct a full test on the whole novels in our dataset because there might be cases of characters annotated as 龙套(the crowds) even if there's a character name in the paragraph or characters that are only annotated with one name but has multiple

[3] https://huggingface.co/uer/roberta-base-chinese-extractive-qa

[4] https://github.com/IDEA-CCNL/Fengshenbang-LM

[5] https://github.com/shtoshni/fast-coref

[6] https://huggingface.co/schen/longformer-chinese-base-4096

[7] https://github.com/fxsjy/jieba

[8] https://huggingface.co/bert-base-chinese

Table 8: Results of classifying the crowds, soliloquy, and dialogues

Method	Accuracy
Rule-based Method (Only for the crowds)	0.6308
BERT classification (Only for the crowds)	**0.8458**
BERT classification	0.72

co-references, so it will take a large effort to re-annotate the novels. But it can be inferred that by using a combination of these models, we can do quote attribution in literary works.

There are still certain challenges that need to be addressed to improve accuracy and robustness further. These challenges include:

1. Longer context for accuracy and co-reference: The pipeline demonstrates a performance drop when dealing with longer contexts, as highlighted in Table 5. This drop is partly attributed to the base model's performance limitations. Additionally, resolving co-reference for names that span across long paragraphs or even chapters, such as "李玉瑶" (Li, Yuyao) and "小瑶" (Xiao Yao), remains unexplored. Instances like these are prevalent in many novels, presenting a complex challenge for accurate co-reference resolution.

2. Eliminating a long chain of models: Due to the diverse range of cases involved in quote attribution, including regular names, sound effects, and crowds, our current approach relies on a long chain of models. However, this has the drawback of previous incorrect predictions affecting subsequent ones. For instance, if the initial BERT classification for crowds and dialogues yields incorrect predictions, subsequent extractive QA processes will also be influenced by these erroneous predictions.

Efforts should be focused on resolving these issues to achieve higher accuracy and enhance the robustness of the quote attribution system.

Limitations

We neglect edge cases for irregular speech content without quotes in Chinese novels in our research. For audio effects, since they only occupy a small portion of the whole novel, we only exclude them by simply defining rules, while there are still a lot of times the rules do not apply.

Our study is only done in modern Chinese literature works. Though the proposed method may be applied to other languages, there might be some language differences that should be taken into account.

7 Future Work

We recognize this work as a stepping stone towards a more comprehensive solution. Here are some promising avenues for further exploration within the domain of automated quote attribution in literary works:

1. Building a Robust Annotated Dataset: A key focus for future work will be the development of a comprehensive and well-annotated dataset specifically designed for quote attribution tasks in fiction. This dataset should encompass a diverse range of writing styles, genres, and complexities to ensure the model generalizes well to unseen data.

2. Unveiling the Potential of Large Language Models (LLMs): LLMs, with their advanced capabilities, including longer context handling and superior understanding, in natural language processing, hold immense potential for quote attribution. Future research will involve exploring the integration of LLMs with quote attribution, potentially leading to a more direct and more accurate result of speaker identification without the combination of multiple models as proposed in this paper. Additionally, the ability of LLMs to parse results in user-defined formats can be a valuable asset, allowing researchers to tailor the output to their specific needs.

These future directions have the potential to significantly improve the accuracy and efficiency of quote attribution in literary analysis. By continuously refining the methodology and exploring new avenues, we can pave the way for a fully automated quote attribution system that empowers researchers and enriches our understanding of literary works.

8 Conclusion

This paper explored the application of machine learning for quote attribution in literary works. By leveraging AI-powered algorithms, we aim to empower literature annotators with faster and more accurate identification of quoted speech sources, ultimately enhancing analysis of fictional works. While writing styles may vary across novels, a significant portion of literary works can benefit from this approach.

References

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

David Elson and Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pages 1013–1019.

Markus Krug. 2020. *Techniques for the Automatic Extraction of Character Networks in German Historic Novels*. Bayerische Julius-Maximilians-Universitaet Wuerzburg (Germany).

Vincent Labatut and Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys (CSUR)*, 52(5):1–40.

Grace Muzny, Michael Fang, Angel Chang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 460–470.

Tim O'Keefe, Silvia Pareti, James R Curran, Irena Koprinska, and Matthew Honnibal. 2012. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799.

Junjie Pan, Lin Wu, Xiang Yin, Pengfei Wu, Chenchang Xu, and Zejun Ma. 2021. A chapter-wise understanding system for text-to-speech in chinese novels. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6069–6073. IEEE.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020a. Learning to ignore: Long document coreference with bounded memory neural networks. *arXiv preprint arXiv:2010.02807*.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020b. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *EMNLP*.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On Generalization in Coreference Resolution. In *CRAC (EMNLP)*.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. *arXiv preprint arXiv:2307.03734*.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Michael Miller Yoder, Sopan Khosla, Qinlan Shen, Aakanksha Naik, Huiming Jin, Hariharan Muralidharan, and Carolyn P Rosé. 2021. Fanfictionnlp: A text processing pipeline for fanfiction. In *The 3rd Workshop on Narrative Understanding*.

Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. 2019. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241.