# SIG: Speaker Identification in Literature via Prompt-Based Generation

**Zhenlin Su**[*1], **Liyan Xu**[†2], **Jin Xu**[†1,3], **Jiangnan Li**[4], **Mingdu Huangfu**[1]

[1]School of Future Technology, South China University of Technology
[2]WeChat AI, Tencent
[3]Pazhou Lab, Guangzhou
[4]Institute of Information Engineering, Chinese Academy of Sciences
zhenlinsu75@gmail.com, liyanlxu@tencent.com, jinxu@scut.edu.cn

## Abstract

Identifying speakers of quotations in narratives is an important task in literary analysis, with challenging scenarios including the out-of-domain inference for unseen speakers, and non-explicit cases where there are no speaker mentions in surrounding context. In this work, we propose a simple and effective approach SIG, a generation-based method that verbalizes the task and quotation input based on designed prompt templates, which also enables easy integration of other auxiliary tasks that further bolster the speaker identification performance. The prediction can either come from direct generation by the model, or be determined by the highest generation probability of each speaker candidate. Based on our approach design, SIG supports out-of-domain evaluation, and achieves open-world classification paradigm that is able to accept any forms of candidate input. We perform both cross-domain evaluation and in-domain evaluation on PDNC, the largest dataset of this task, where empirical results suggest that SIG outperforms previous baselines of complicated designs, as well as the zero-shot ChatGPT, especially excelling at those hard non-explicit scenarios by up to 17% improvement. Additional experiments on another dataset WP further corroborate the efficacy of SIG.

## 1 Introduction

Speaker identification in literary text aims at identifying the speaker of quotation in narrative genres such as fictions or novels (Elson and McKeown 2010), serving as an important step for downstream applications such as novel-to-script conversion (Soo, Yang, and Soo 2019).

Table 1 shows the examples of three types of quotation in the PDNC dataset proposed for this task (Vishnubhotla, Hammond, and Hirst 2022), based on whether the speaker is explicitly indicated by an adjoining expression (explicit), or appears without an attribution (implicit), or is indicated by an anaphoric mention (anaphoric). In this paper, the latter two are referred to as the "non-explicit" case that cannot be solved by simple superficial narrative patterns, calling for a deep global understanding of the context surrounding the quotation, of which the importance is also highlighted by other related tasks, such as character comprehension in narrative stories (Yu et al. 2022; Sang et al. 2022). It becomes even more challenging for cross-domain inference, where the system neither has prior knowledge about candidate speakers, nor it could identify the absent speaker mentioned in context for those implicit cases. Ideally, a practical system should be able to determine the speakers under any circumstances, achieving an open-world speaker identification paradigm.

| **Quote Type:** Explicit |
|---|
| **Quotation** (w/ context):<br>Mrs. Elton hardly waited for the affirmative.<br>**"Well, we shall see."** said Mrs Elton.<br>Emma was almost too much astonished to answer<br>**Speaker**: Mrs Elton |
| **Quote Type:** Anaphoric |
| **Quotation** (w/ context):<br>But Mrs. Elton was very much discomposed indeed.<br>She said:**"Rather he than I!"**,<br>**Speaker**: Mrs. Elton |
| **Quote Type:** Implicit |
| **Quotation** (w/ context):<br>"Your father will not be easy; why do not you go?"<br>**"I am ready, if the others are."**<br>"Shall I ring the bell?"<br>**Speaker**: Emma |

Table 1: Examples of three quote types in PDNC. The quotation in each example is highlighted in bold.

Early works on this task involve linguistic-based properties and designs that can be effective especially for explicit cases (Elson and McKeown 2010). With the remarkable success of pretrained language models (PLMs) such as BERT (Devlin et al. 2019), recent studies have shifted their focus towards leveraging PLMs for speaker identification (Liu et al. 2019), (Pan et al. 2021). However, several problems still remain. On the one hand, certain previous works divide the speaker identification task into interrelated subtasks executed with separately trained models, including named entity recognition and coreference resolution ((Pan et al.

[*]Primary work done in WeChat AI.
[†]Co-corresponding authors.

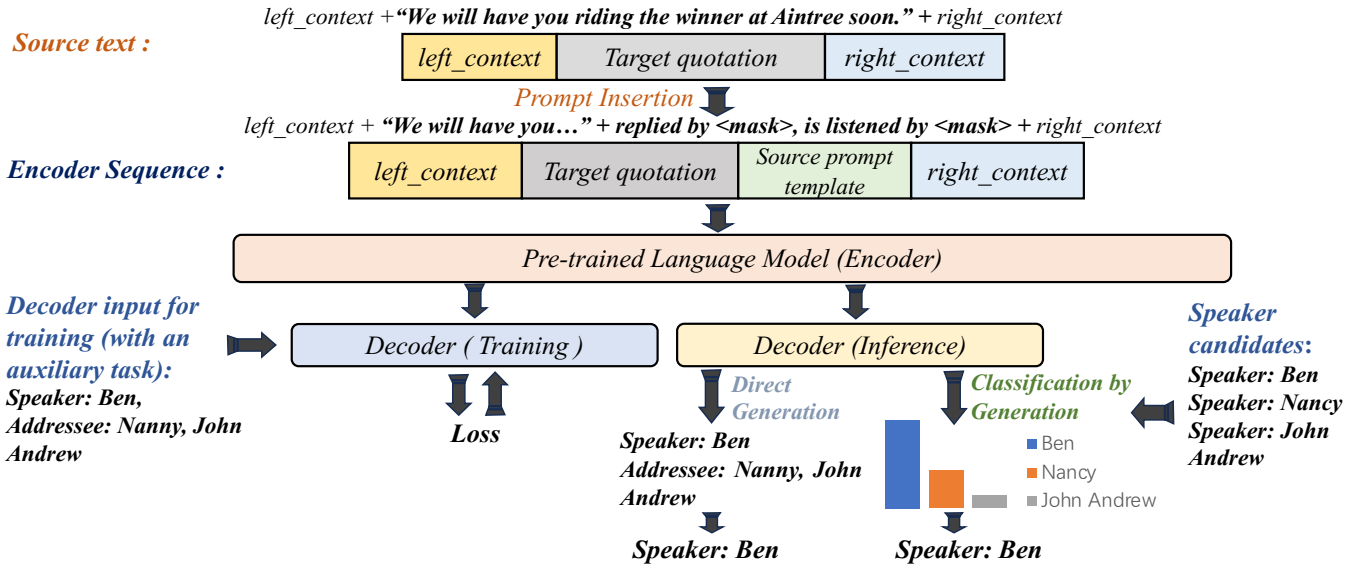



Figure 1: Illustration of our proposed approach SIG. Encoder input and decoder output are formatted according to the designed prompt templates described in Section 3.2. During inference, SIG either generates the speaker directly or determines the speaker based on the highest generation probability of each candidate.

2021), Yoder et al. (2021)). These approaches face inherent challenges due to the inevitable error propagation from sub-tasks that undermines the final performance (Yu, Zhou, and Yu 2022a). For instance, the performance of the recent state-of-the-art coreference resolution model is 85.09 F1 (Miculicich and Henderson 2022) on the standard CoNLL-12 benchmark (Pradhan et al. 2012), from which the wrong entity clusters will be accumulated to the next step. On the other hand, other previous works regard speaker identification as span extraction to extract the speaker mention spans directly from the input (Yu, Zhou, and Yu 2022a), which may fail in the implicit cases where such mention spans can not be found in the context surrounding the quotation. For anaphoric cases, these approaches are sometimes only able to capture pronouns without localizing the explicit speaker identity. Additionally, when dealing with implicit instances, they will always produce incorrect speaker spans, due to the absence of a correct answer to extract.

Recently, prompt-based methods have been highlighted by generative Large Language Models (LLMs) such as ChatGPT. To evaluate LLM's capability on this task, we also conduct experiments with zero-shot prompting. Nevertheless, inspired by combining prompt engineering and generative methods, in this work, we propose a new approach for the open-world speaker identification paradigm, which utilizes generation models and converts the task as open-world classification, choosing the speaker according to generation probability, and being able to adopt labels in any form.

Specifically, our approach dubbed SIG (**S**peaker **I**dentification via **G**eneration), encodes the quotation and its context as the input based on our designed prompt template, then it can either opt to generate the speaker directly, or to evaluate the generation probability for any speaker candidates. For the latter case, during inference, possible candidate speakers are enumerated, and the speaker with the highest generation probability to fill in the prompt template is selected as the final answer. In addition, SIG also introduces auxiliary tasks to capture more implicit information related to speaker identification, such as predicting addressees, which is seamlessly integrated into the designed prompt template as well.

Compared to previous methods, SIG can offer notable advantages. Firstly, SIG approaches the speaker identification task in an end-to-end manner, bridging the input and output through the prompt template by generation, which circumvents the inherent drawbacks associated with traditional pipeline approaches comprising multiple sub-tasks. Secondly, it is more flexible to handle a variety of situations, including the implicit cases where the correct answer does not appear in the context, as well as the cross-domain cases where the speaker is not seen by the model during training.

Our empirical experiments are mainly conducted on the Project Dialogism Novel Corpus (PDNC) (Vishnubhotla, Hammond, and Hirst 2022), of which about 66% are non-explicit cases. The results demonstrate that SIG outperforms the state-of-the-art approaches on PDNC by a significant margin of 4% averaged accuracy for all quotations and a substantial 17% averaged accuracy specifically for non-explicit cases. In addition, ChatGPT presents strong performance; our approach still outperforms zero-shot ChatGPT (GPT-3.5-turbo) by 9% for all quotations and 2% for non-explicit quotations. These results indicate that our method achieves state-of-the-art performance on the cross-domain speaker identification task on PDNC, especially for complex cases. In addition, SIG is also evaluated on WP (Chen, Ling, and Dai 2019), a speaker identification dataset stemmed from a Chinese novel, and shown surpassing previous baselines by 5.2% on the test set. It is worth noting that although this

work does not use LLMs for training due to the computational resource constraints, our proposed approach could adopt LLMs in the future for further enhancement.

Overall, our contributions to this work are fourfold:

- We propose SIG, an approach for speaker identification via prompt-based generation, achieving open-world classification on speakers, to be a new paradigm for this task.
- Experiments suggest that SIG outperforms existing methods as well as zero-shot ChatGPT on PDNC, demonstrating superior performance on cross-domain evaluation and the harder non-explicit cases.
- SIG is also shown to outperform baselines on WP for the in-domain evaluation setting, showing the robust performance of our proposed approach.

## 2 Background and Related Work

**Speaker Identification** Previous works have proposed several datasets for evaluation. Elson and McKeown (2010) introduces the CQSA corpus, which contains quotations from 4 novels and 7 short stories that are annotated. Bamman, Lewke, and Mansoor (2020) released LitBank corpus, an annotated dataset of 100 English fictions, aimed at supporting various tasks including entity recognition, coreference resolution, and speaker attribution. The recently proposed Project Dialogism Novel Corpus (PDNC) (Vishnubhotla, Hammond, and Hirst 2022) comprises 22 comprehensive works of fiction, encompassing 12,773 explicit quotations and 23,205 non-explicit quotations, being the largest dataset for training and evaluation.

**Previous Approaches** Elson and McKeown (2010) first introduces quote attribution in literary narratives by assigning speaker tags to quoted speech. They utilize rule-based and statistical learning techniques to identify candidate characters, determine their genders, and attribute each quote to the most likely speaker. Muzny et al. (2017) proposes a deterministic, two-step methodology for quotation attribution. A sequence of progressively intricate filters is utilized to initially associate each quotation with a mention, and then it proceeds to link the mention to a character entity.

Moving to the era of PLMs, Chen, Ling, and Liu (2021a) proposes a candidate scoring network based on BERT along with a revision algorithm to identify the speaker. Yu, Zhou, and Yu (2022b) resolves speaker identification as a span extraction task and applies the extractive Question-Answering paradigm to address it. Pipeline approaches are also mainstream (Bamman, Lewke, and Mansoor 2020; Xu and Choi 2022), utilizing BERT to independently train base models for modules such as coreference resolution (Xu and Choi 2020). Vishnubhotla, Hammond, and Hirst (2022) enhances this approach by restricting the set of candidates to resolved mention spans from the coreference resolution step for direct quotation-to-entity linking, resulting in state-of-the-art performance on PDNC. These studies highlight the significance of PLMs for speaker identification while also leaving room for improvement due to several disadvantages as discussed above.

**Template-based Approach** Brown et al. (2020) is among the first works to utilize prompts for solving text classification tasks. Schick and Schütze (2020) further employs a template-based approach to transform text classification into a cloze-style problem, by training the model to fill in the provided slots. Cui et al. (2021) has formulated templates for both input and output, selected the optimal choice by the option of the most likely generation, addressing the few-shot Named Entity Recognition (NER). The generative paradigm has also shown success in other NLP tasks (Xu et al. 2017; Lu et al. 2021). In this work, we adopt the generative paradigm for open-world classification that has been advocated in other areas (Xu et al. 2023). The success of these approaches in other areas demonstrates the feasibility of incorporating specific prompt templates for task verbalization aligned with the pretraining stage.

## 3 Methodology

### 3.1 Task Definition

Before delving into the specifics of our proposed approach, the basics of the speaker identification task are clarified as follows: Within the context of narrative corpus, sentences can be categorized into two distinct types: those containing direct speech by characters (referred to as quotations), and those comprising narrative descriptions. Our task revolves the extraction of pertinent information from the contextual surroundings of quotations that is subsequently employed to determine the most likely speaker. The names and aliases of potential speakers can be pre-collected. Our approach targets to identify the speaker for a given quotation accompanied by its surrounding context.

### 3.2 Approach Introduction

Figure 1 illustrates our proposed approach SIG. It takes the quotation along with its left and right context as template-input, generating the speaker directly, or a score for each candidate speaker. To initiate the process, the source prompt template is inserted with a placeholder <mask> positioned after the quotation. Guided by the target prompt template, the model encodes the designed input and starts the auto-regressive mechanism for generation to fill the speaker candidate in the target template.

As shown in Figure 1, SIG inserts the prompt containing the placeholder after the quotation, and defines the task flexibly verbalized by natural language, which minimizes the gap between the pretraining and finetuning for models such as BART (Lewis et al. 2020), and helps to better leverage the internal knowledge of PLMs.

**Prompt Template Design** Different prompt templates are examined as shown in Table 3. Concretely, both the source (input) and target (output) can be augmented with prompts such as "*replied by:*" or "*speakers:*". For the source template, it is followed by <mask>, processed by the encoder; while for the target, the decoder is expected to fill with the correct speaker. Additionally, we also experiment with the naive version without any templates, where the decoder generates the speaker directly based on the encoded quotation and context, without using the placeholder <mask>. These prompts verbalize the task naturally and instruct the model to pay more attention towards the quotation.

**Prompt Template with Auxiliary Task** One of the advantages for prompt-based task verbalization is that we can easily integrate other tasks together by adding more instructions to the template. In this work, we also investigate incorporating the auxiliary task for speaker identification. Here, we train the model to simultaneously recognize both the speaker and the addressees of a quotation, learning to disentangle the multi-party interactions in a dialogue, which may potentially bolster the quotation attribution itself. Moreover, given that both the addressees and the speaker are typically represented by personal names, predicting the addressees strengthens the model's attention towards the pertinent person names in the context.

In our approach to integrate with the auxiliary addressee prediction, we simply extend the prompt by adding "*is listened by* <mask>" in the source template, and "*Addressee:*" in the target template, as depicted in Figure 1. For the inference, the model will still produce the speaker first, and the predicted addressees could be discarded.

Table 4 also shows results with other auxiliary tasks, e.g. gender identification. SIG adopts addressee prediction as it leads to the best result empirically.

### 3.3 Inference

**Direct Generation** As SIG operates upon the generative model, it is straightforward to generate the speaker directly. We take the quotation and its context as the input, formatted by the prompt template, and model generates the speaker of the quotation without decoding constraints. The predicted speaker is then parsed from the generated output. For evaluation, it is judged whether it belongs to one of the speaker candidates or aliases from gold labels. However, this process could potentially introduce the following disadvantages. First, the generated output sometimes does not match the corresponding speaker accurately. For example, if the string generated by the model is *Beaver*, it is difficult to resolve whether it refers to *Mrs. Beaver* or *Mr. Beaver*. Second, it is harder to control the generation, especially when the model also adopts other auxiliary tasks.

**Classification by Generation** To address the drawbacks in direction generation, we propose to guide the generation process by providing the speaker candidates to the decoder. We begin by listing all candidate speakers for the given quotation, and for each candidate, we obtain its generation probability according to the trained model. Thus, this paradigm supports any forms of speakers, including new speakers unseen during training, or speakers that do not appear in the surrounding context.

As speaker names can often be of different lengths, in order to cope with this situation, we take the averaged probability of the output, and select the one with the highest probability to be the final prediction. The score for the $i$th candidate can be denoted as follows:

$$f(\mathbf{T}_i) = \sum_{c=1}^{m} p(t_c|t_{1:c-1}, \mathbf{X})/m, \tag{1}$$

where $T_i = (t_1, .., t_m)$ is the target template output with length $m$, $X$ is the source input, and $p$ represents the probability from the decoder at each step.

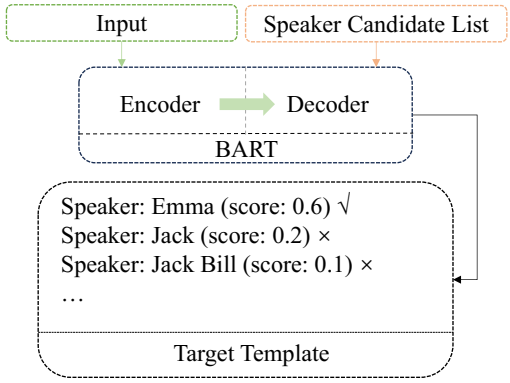



Figure 2: Classification by generation: each speaker candidate is fed to the decoder, and its generation probability is obtained; the highest option is selected as the final prediction.

### 3.4 Training

Given a speaker quotation $\mathbf{q}$, its left context $\mathbf{c}^l$ and right context $\mathbf{c}^r$, the source prompt template $\mathbf{t_s}$, the input for the model $\mathbf{X}$ is formulated as follows:

$$[\text{CLS}] \oplus \mathbf{c^l} \oplus \mathbf{q} \oplus \mathbf{t_s} \oplus \mathbf{c^r} \oplus [\text{SEP}], \tag{2}$$

where $\oplus$ represents concatenation (example in Figure 1). The model encodes the provided input, and the output is conditioned on the hidden state of the input, following the standard encoder-decoder architecture. The training adopts teacher-forcing, where the model is optimized to minimize the negative log-likelihood of the gold output sequence, denoted as below:

$$\mathcal{L} = -\sum_{c=1}^{m} \log p(t_c|t_{1:c-1}, \mathbf{X}) \tag{3}$$

The training is the same for either direction generation or classification by generation. Thus, the trained model is flexible to choose the inference according to specific scenarios.

## 4 Cross-Domain Experiments

### 4.1 Dataset

Our main experiments are conducted on Project Dialogism Novel Corpus (PDNC) (Vishnubhotla, Hammond, and Hirst 2022), a recently introduced dataset specifically designed for analysis on English literary text. PDNC stands out as the largest dataset for this task, encompassing a diverse range of genres such as science fiction, literary fiction, children's literature and detective fiction, each contributing to a stylistically varied speech style.

To elaborate, PDNC consists of a compilation of 22 extensive novels, comprising a total of 35,978 identified and annotated quotations that enable training. These annotations capture crucial attributes including the speaker, addressees,

and quotation type. Quotations within PDNC are categorized into three distinct types as introduced in Table 1: explicit, implicit, anaphoric.

## 4.2 Evaluation Protocol

As PDNC consists of multiple novels, it enables cross-domain evaluation, such that the model is evaluated on the test set that has no overlapped novels from the training. It is a more practice scenario, which is also adopted by PDNC authors (Vishnubhotla, Hammond, and Hirst 2022; Vishnubhotla et al. 2023), as the trained model should be able to recognize speakers upon any literary, rather than only for those limited novels seen during training.

Following the evaluation outlined by previous works, our results are reported as averaged accuracy over five experiments. For each experiment, four novels are randomly selected as the test set, while the remaining novels are used for training the model. The process ensures that no novels are select twice in the test set.

In addition, previous works also exclude instances from training and evaluation where speakers have fewer than 10 annotated quotations. This step was taken to mitigate the potential impact of minor characters, which often fall under the long-tail distribution, thereby ensuring a more focused and insightful analysis. Our experiments keep the same setting, to be comparable with previous approaches.

## 4.3 Experimental Approaches

**BookNLP** BookNLP is an NLP tool tailored for English literary text. Its processing pipeline encompasses various tasks, including but not limited to named entity recognition, conference resolution, and speaker attribution. During the speaker attribution phase, BookNLP initiates the process by employing the entity recognition and conference resolution models to identify person clusters within a specified window surrounding the quotation. A BERT model incorporating contextual, positional, and gender cues then assigns scores to mention spans located within this window. The mention span with the highest score is then selected as the attributed speaker.

To ensure alignment between the extracted mention spans from BookNLP and the labels within the PDNC corpus, we suitably relax the matching criteria. Any mention span that corresponds to a mention span within the PDNC is considered correct. Additionally, if an incorrect mention span is clustered to the correct answer, it is still considered a valid choice.

**BookNLP+** In a complementary effort, Vishnubhotla et al. (2023) enhances BookNLP by constraining the set of candidates to resolve mention spans stemming from the coreference resolution step. This refined approach directly establishes a link between quotations and entities, achieving state-of-the-art results for the cross-domain evaluation on PDNC.

**Large Language Models (LLMs)** We employ ChatGPT (*gpt-3.5-turbo-0613*) from OpenAI for the zero-shot experiments. Certain efforts for prompt engineering are performed to refine the task prompts. Especially, Chain-of-Thought (CoT) (Wei et al. 2022) is also adopted for all LLM experiments. Due to the expense of API usage, we conduct only a single evaluation of ChatGPT.

Since LLMs are not trained for this task, we employ a lenient metric for evaluation. A response is considered correct as long as the true speaker's name (or one of its aliases) appears as a substring in the response. Thus, our reported results for ChatGPT can be considered as an upper bound performance under our specific prompt setting.

| | Non-Explicit | Total |
|---|---|---|
| BookNLP | 0.46/0.39 | 0.68/0.66 |
| BookNLP+ | 0.53*/- | 0.68*/- |
| ChatGPT | 0.70/0.71 | 0.71/0.71 |
| $SIG^D$ | 0.56/0.51 | 0.57/0.54 |
| SIG | **0.70/0.73** | **0.72/0.78** |

Table 2: Accuracy for the cross-domain evaluation on PDNC, broken down by quotation types. For BookNLP+, we take the reported accuracy from Vishnubhotla et al. (2023). For each other approach, we report both the mean and median accuracy of the five repeated experiments described in Section 4.2. Our proposed approach, SIG is shown to obtain the best accuracy.

**RoBERTa** We also adopt RoBERTa (Liu et al. 2019) as the encoder-only model for the conventional classification paradigm, where a linear layer is stacked on the last layer of hidden state, followed by softmax to classify speakers directly. We keep the same prompt template as SIG for the source input. However, since this method can only perform in-domain evaluation, and cannot handle speakers not seen during training, we report the in-domain evaluation results separately from the main cross-domain evaluation.

**$SIG^D$** This method is the **D**irect generation setting of our proposed approach, described in Section 3.3.

**SIG** This method adopts the classification as generation setting of our approach, which is more flexible than $SIG^D$. Both $SIG^D$ and SIG employ BART (Lewis et al. 2020) as the sequence generation PLM.

## 4.4 Results and Discussions

Table 2 shows the cross-domain evaluation results with approaches described in Section 4.3. Several observations could be made as follows.

First, Table 2 suggests that SIG surpasses all baselines, including the zero-shot ChatGPT, for both non-explicit quotations or for all quotation types, achieving the new state-of-the-art for the cross-domain setting. Notably, our proposed method demonstrates an improvement of 4% for all types, and especially a significant 17% for the non-explicit quotations. The superior results underscore SIG's robust generalization capability to identify speakers in unseen domains and novels, as well as recognizing the speakers for non-explicit quotations beyond the superficial cues.

| Source Template | Target Template | Accuracy |
|---|---|---|
| No Source template | No target template | 56.34 |
| "Quotation + replied by: <mask>" | No target template | 58.12 |
| "Quotation + replied by: <mask>" | replied by: <Candidate_speaker> | 66.32 |
| "**Quotation + replied by: <mask>**" | **Speaker: <Candidate_speaker>** | **68.43** |
| "Quotation + Speaker: <mask>" | replied by: <Candidate_speaker> | 58.14 |
| "Quotation + Speaker: <mask>" | Speaker: <Candidate_speaker> | 64.44 |

Table 3: Results on PDNC using different prompt templates with BART, described in Section 3.2. Though these prompts are quite simple, evaluation suggests that they indeed have quite an impact on the final performance. SIG adopts the prompt template of the best performance highlighted by bold.

Second, ChatGPT obtains impressive results, despite it has not undergone any task-specific training for PDNC. Particularly, Table 2 shows that by simply performing zero-shot inference along with appropriate prompt engineering, ChatGPT outperforms both BookNLP and BookNLP+ that are of sophisticated approach design by good margins. It indicates that LLM-based methods are future-proofing, such that they either resolve tasks directly through their powerful understanding and reasoning abilities, or can be used for training that serves as a direct substitution of conventional generative models such as BART. In this work, we do not adopt training with open-source LLMs due to constraints of our computational resources.

### 4.5 Ablation Studies

**Prompt Templates** Table 3 shows evaluation results with various prompt templates for speaker identification without adding auxiliary tasks. Notably, by judiciously selecting an appropriate pair of prompt templates, the performance receives a substantial increase of **12.1%** compared to not using any prompts at all. In light of these outcomes, SIG adopts the best-performing template from these subsequent experiments. The results further corroborates the importance to align the finetuning task to the model's pretraining stage to better induce its internal knowledge.

**Auxiliary Tasks** The performance of speaker identification accompanied by different auxiliary tasks is depicted in Table 4. As observed, training with either addressee identification (71.59%) or gender identification (69.51%) yields enhanced results compared to training without any auxiliary tasks (68.43%). Conversely, when the model is trained with fiction classification to identify which novels the quotation comes from, there is a noticeable decline by more than 5%. These findings underscore the notion that not all auxiliary tasks are able to boost the main task of speaker identification. Consequently, a careful selection of appropriate auxiliary tasks is imperative to optimize performance.

| Auxiliary Task | Accuracy |
|---|---|
| None | 68.43 |
| Fiction Identification | 63.38 |
| Gender Identification | 69.51 |
| **Addressee Identification** | **71.59** |

Table 4: Results on PDNC adopting different auxiliary tasks. SIG employs the addressee prediction as an auxiliary task, as it achieves the best performance.

| | Train | Test |
|---|---|---|
| Cross-Domain | 28105 (32.8%) | 5989 (34.7%) |
| In-Domain | 11235 (100%) | 22589 (0%) |

Table 5: The number of quotations of the training and test set on PDNC for the cross-domain evaluation and in-domain evaluation. The ratio of explicit quotations is shown in parentheses.

## 5 In-Domain Evaluation

### 5.1 PDNC Experiments

Distinct from the cross-domain evaluation in Section 4, in-domain evaluation provides the same set of novels for both training and evaluation. Thus, speaker candidates remain the same, and the trained model gains prior knowledge regarding those speakers after finetuning.

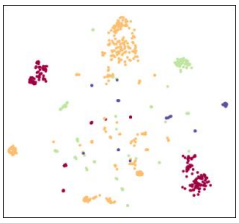
Though the in-domain evaluation is a less practical setting, as the model may not be able to recognize unseen speakers, it can be regarded as an upper bound of the model capability, since for the same novel, a model with prior speaker knowledge is likely to outperform the model that is trained on other novels. For this setting, we compare SIG with the conventional classification paradigm, where we utilize RoBERTa, described in Section 4.3, to directly classify the attributed speaker for a quotation during evaluation.

Following the setting outlined in (Vishnubhotla, Hammond, and Hirst 2022), the training and test set for each novel is organized based on quotation types. Specifically, the training set comprises only explicit quotations, while the remaining quotations (implicit and anaphoric) are assigned to the test set. This is a harder setting than randomly splitting quotations as training or evaluation, as now the model does not see any non-explicit quotations during training, which requires generalization of the model capacity. The overall statistics for cross-domain and in-domain evaluation on PDNC is provided in Table 5.

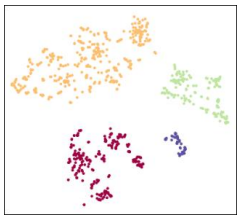
Table 6 shows the evaluation results by top-k accuracy

| | Top 1-5 Accuracy |
|---|---|
| RoBERTa | **69.39** / **78.56** / **80.23** / 82.45 / 84.55 |
| SIG | 64.45 / 75.32 / 78.36 / **83.16** / **86.31** |

Table 6: The performance for the in-domain speaker identification. Top 1-5 refers to the correct answer being within the top 1 to top 5 predictions by the model.



(a) Output embedding from RoBERTa.



(b) Speaker embedding from SIG.

Figure 3: t-SNE visualization of the embedding distribution on the test set for PDNC. Output from the same novel is marked by the same color.

of RoBERTa and SIG. Furthermore, Figure 3 provides the t-SNE visualization of the output embedding from both approaches. For RoBERTa, it is the averaged embedding of a quotation that is used for classification; for SIG, it is the decoder embedding of the speaker names.

Upon analyzing the results in Table 6 and Figure 3, we could arrive at the following conclusions:

- Table 6 shows that even compared to RoBERTa which is designed specifically for in-domain speaker identification, SIG lags behind by less than 5% without any modification, and starts to surpass Roberta consistently when $k \geq 4$ (by 0.7% when k = 4, with larger margin when k is larger). Importantly, SIG is also ready to perform cross-domain evaluations that RoBERTa cannot handle at all.
- By Figure 3, it becomes apparent that the output embedding of SIG exhibits clustering based on novels, whereas certain embedding of RoBERTa is notably distant from the clusters. This observation suggests that SIG is adept at capturing the relationships associated with the labels of candidate speakers.

### 5.2 WP Experiments

WP (Chen, Ling, and Dai 2019) is a dataset annotated on the Chinese novel *World of Plainness*. The name list was collected manually and contained 125 roles that occurred throughout the novel. Chen, Ling, and Liu (2021b) further extended this dataset by making additional annotations and obtained 2596 quotation instances in total. WP only supports in-domain evaluation as it only comprises one novel.

**Baselines**

- Candidate scoring network (CSN) (Chen, Ling, and Liu 2021b): this method follows a three-step pipeline, nearest mention location(NML), candidate scoring network(CSN) and speaker alternation pattern. Initially, an input instance accompanied by the name list is sent to NML to obtain a set of speaker candidates. Subsequently, each candidate is sent to CSN to generate its score. A revision based on SAP is then implemented on the quotation that produces the final decision.
- End-to-End Speaker Identification (E2E-SI) (Yu, Zhou, and Yu 2022b): for each quotation input, this method extracts speaker spans appeared in surrounding context, similar to the extractive Question-Answering model, where it identifies start and end positions of the predicted speaker spans for the final decision.

Since WP does not provide additional information such as addressees, all models are trained for speaker identification only, without other auxiliary tasks.

| | Dev | Test |
|---|---|---|
| CSN | - | 82.50 |
| E2E-SI | 78.60 | 80.90 |
| $SIG^D$ | 85.27 | 85.89 |
| SIG | **85.81** | **86.15** |

Table 7: Accuracy for the in-domain evaluation on WP. Baselines are described in Section 5.2, and their results are directly taken from the original papers.

**Results** Table 7 shows the evaluation results on WP. SIG outperforms the two baselines by 3.6% and 5.2% respectively. Particularly, $SIG^D$ obtains comparable performance as SIG, indicating that even with direct generation without seeing the speaker candidates, the simple generation paradigm with prompt templates could still be superior to the pipeline or encoder-only approaches.

## 6 Conclusions

In this work, we propose SIG, an approach that supports the open-world speaker identification on literary text. SIG adopts the generation-based method, verbalizing the task and quotation input based on designed prompt templates; especially, SIG is flexible integrating other auxiliary tasks by simply extending the prompt. The inference can be either direct generation, or select the speaker candidate with the highest generation probability, enabling inference on different domains whose speakers are not seen by the model during training. Evaluation on PDNC demonstrates that SIG surpasses all baselines, as well as the zero-shot ChatGPT, particularly excelling in cross-domain non-explicit speaker identification scenarios that demand a profound comprehension of context. Additional in-domain evaluation on PDNC and WP further confirms the efficacy of SIG.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (62372187) and Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

## References


Bamman, D.; Lewke, O.; and Mansoor, A. 2020. An Annotated Dataset of Coreference in English Literature. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, 44–54. European Language Resources Association.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chen, J.-X.; Ling, Z.-H.; and Dai, L.-R. 2019. A Chinese Dataset for Identifying Speakers in Novels. In *INTERSPEECH*, 1561–1565. Graz, Austria.

Chen, Y.; Ling, Z.; and Liu, Q. 2021a. A Neural-Network-Based Approach to Identifying Speakers in Novels. In Hermansky, H.; Cernocký, H.; Burget, L.; Lamel, L.; Scharenborg, O.; and Motlícek, P., eds., *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, 4114–4118. ISCA.

Chen, Y.; Ling, Z.-H.; and Liu, Q.-F. 2021b. A Neural-Network-Based Approach to Identifying Speakers in Novels. In *Interspeech*, 4114–4118.

Cui, L.; Wu, Y.; Liu, J.; Yang, S.; and Zhang, Y. 2021. Template-Based Named Entity Recognition Using BART. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 1835–1845. Association for Computational Linguistics.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Elson, D. K.; and McKeown, K. R. 2010. Automatic Attribution of Quoted Speech in Literary Narrative. In Fox, M.; and Poole, D., eds., *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, Y.; Lin, H.; Xu, J.; Han, X.; Tang, J.; Li, A.; Sun, L.; Liao, M.; and Chen, S. 2021. Text2Event: Controllable Sequence-to-Structure Generation for End-to-end Event Extraction. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2795–2806. Online: Association for Computational Linguistics.

Miculicich, L.; and Henderson, J. 2022. Graph Refinement for Coreference Resolution. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2732–2742. Dublin, Ireland: Association for Computational Linguistics.

Muzny, G.; Fang, M.; Chang, A. X.; and Jurafsky, D. 2017. A Two-stage Sieve Approach for Quote Attribution. In Lapata, M.; Blunsom, P.; and Koller, A., eds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, 460–470. Association for Computational Linguistics.

Pan, J.; Wu, L.; Yin, X.; Wu, P.; Xu, C.; and Ma, Z. 2021. A Chapter-Wise Understanding System for Text-To-Speech in Chinese Novels. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, 6069–6073. IEEE.

Pradhan, S.; Moschitti, A.; Xue, N.; Uryupina, O.; and Zhang, Y. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40. Jeju Island, Korea: Association for Computational Linguistics.

Sang, Y.; Mou, X.; Yu, M.; Yao, S.; Li, J.; and Stanton, J. 2022. TVShowGuess: Character Comprehension in Stories as Speaker Guessing. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4267–4287. Seattle, United States: Association for Computational Linguistics.

Schick, T.; and Schütze, H. 2020. Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference. *CoRR*, abs/2001.07676.

Soo, M.; Yang, Y.; and Soo, V. 2019. Automatic Conversion of a Chinese Fairy Story into a Script - A Preliminary Report and Proposal. In *2019 International Conference on Technologies and Applications of Artificial Intelligence, TAAI 2019, Kaohsiung, Taiwan, November 21-23, 2019*, 1–6. IEEE.

Vishnubhotla, K.; Hammond, A.; and Hirst, G. 2022. The Project Dialogism Novel Corpus: A Dataset for Quotation Attribution in Literary Texts. In Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; Mazo, H.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, 5838–5848. European Language Resources Association.

Vishnubhotla, K.; Rudzicz, F.; Hirst, G.; and Hammond, A. 2023. Improving Automatic Quotation Attribution in Literary Novels. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 737–746. Association for Computational Linguistics.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 24824–24837. Curran Associates, Inc.

Xu, J.; Tang, B.; He, H.; and Man, H. 2017. Semisupervised Feature Selection Based on Relevance and Redundancy Criteria. *IEEE Transactions on Neural Networks and Learning Systems*, 28(9): 1974–1984.

Xu, L.; and Choi, J. 2022. Modeling Task Interactions in Document-Level Joint Entity and Relation Extraction. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5409–5416. Seattle, United States: Association for Computational Linguistics.

Xu, L.; and Choi, J. D. 2020. Revealing the Myth of Higher-Order Inference in Coreference Resolution. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8527–8533. Online: Association for Computational Linguistics.

Xu, L.; Zhang, C.; Li, X.; Shang, J.; and Choi, J. D. 2023. Towards Open-World Product Attribute Mining: A Lightly-Supervised Approach. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12223–12239. Toronto, Canada: Association for Computational Linguistics.

Yoder, M.; Khosla, S.; Shen, Q.; Naik, A.; Jin, H.; Muralidharan, H.; and Rosé, C. 2021. FanfictionNLP: A Text Processing Pipeline for Fanfiction. In *Proceedings of the Third Workshop on Narrative Understanding*, 13–23. Virtual: Association for Computational Linguistics.

Yu, D.; Zhou, B.; and Yu, D. 2022a. End-to-End Chinese Speaker Identification. In Carpuat, M.; de Marneffe, M.; and Ruíz, I. V. M., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, 2274–2285. Association for Computational Linguistics.

Yu, D.; Zhou, B.; and Yu, D. 2022b. End-to-End Chinese Speaker Identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2274–2285. Seattle, United States: Association for Computational Linguistics.

Yu, M.; Sang, Y.; Pu, K.; Wei, Z.; Wang, H.; Li, J.; Yu, Y.; and Zhou, J. 2022. Few-Shot Character Understanding in Movies as an Assessment to Meta-Learning of Theory-of-Mind. arXiv:2211.04684.