Advances and Challenges in the Automatic Identification of Indirect Quotations in Scholarly Texts and Literary Works

Frederik Arnold, Robert Jäschke, Philip Kraut
Humboldt-Universität zu Berlin
{frederik.arnold, robert.jaeschke, ph.kraut}@hu-berlin.de

Abstract

Literary scholars commonly refer to the interpreted literary work using various types of quotations. Two main categories are direct and indirect quotations. In this work we focus on the automatic identification of two subtypes of indirect quotations: paraphrases and summaries. Our contributions are twofold. First, we present a dataset of scholarly works with annotations of text spans which summarize or paraphrase the interpreted drama and the source of the quotation. Second, we present a two-step approach to solve the task at hand. We found the process of annotating large training corpora very time consuming and therefore leverage GPT-generated summaries to generate training data for our approach.

1 Introduction

Literary scholars reproduce literary works in different ways and have to decide how precise their reference to the interpreted text should be. Direct quotation, using direct speech and quotation marks, is considered the closest, the verbatim rendition of a source. No information is omitted (except the surrounding context and, sometimes, marked or unmarked omissions in the quotation) or added. To a certain degree, direct quotations preserve the poetic form of a text. Retaining the literality of the source and their precise wording is one of the canonical features of the concept quotation (Helmstetter, 2003). Recent literary theory has categorized various types of references to literary texts that are used in scholarly interpretative articles (Winko, 2022).

In our research project *Key passages in literary works*,[1] we use methods of Computational Literary Studies to find intensively interpreted passages. We identify these *key passages* by accumulating direct quotations of a literary text in scholarly texts, which led us to detailed insights into the scholars' quotation practices (Arnold and Jäschke, 2021, 2023). We consider the heavily quoted passages in academic texts as key for the particular exegesis. We recognize that not only direct quotations play an important role in interpretive practices but also indirect quotations. Therefore, in this work, we develop and analyze methods to automatically identify indirect quotations in scholarly texts and literary works.

We follow the definition from Winko (2022): An indirect quotation translates object language into metalanguage without adding essential information that does not stem from the textual source itself. Paraphrases and summaries are subcategories of indirect quotation. A paraphrase is more or less a recurrence of the content with a change of the wording (de Beaugrande and Dressler, 1981), whereas a summary abbreviates the content, with a change of the wording, too.

Indirect quotation is only one of several types of references scholarly interpretations use. In interpretive texts, scholars also apply classification, illustration, explanation, explication, and exegesis (Winko, 2022). All these types of interpretive practices need extrinsic context information whereas types of direct and indirect citation – generally speaking – only use intrinsic features of the literary text. Additionally, they vary significantly from quotations because they include information that comes from the interpreting scholar who writes the interpretative article. These references and quotations are often mixed and distinguishing occurrences of indirect quotations from the surrounding text and differentiating between the distinct types is a hard task, even for human experts.

Direct quotations are easier to identify as they are syntactically marked, for example, by quotations marks, and can be identified and linked using existing tools, such as Quid and ProQuo (Arnold and Jäschke, 2021, 2023). Indirect quotations, on the other hand, are much more challenging. Often they are not accompanied by any surface indicator and therefore we do not have prior knowledge of the location of candidates in a scholarly work. Sometimes, scholars mention the source of an indirect quotation in the running text or in a footnote. However, these references are applied rather non-systematically and cannot reliably be utilized. Additionally, the length of indirect quotations can vary from very short – only a couple of words – to full, or even multiple, sentences.

Another big challenge is the non-existence of annotated training data and we found that annotating this phenomenon is a very time-consuming process and an arduous task for human annotators.

Considering these challenges, we made the following decisions. First, we want to avoid manually creating large corpora for training machine learning models. Sec-

[1] https://hu.berlin/keypassages

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 179–190

ond, we focus on dramas which are available in cleaned and annotated form from DraCor (Fischer et al., 2019). This allows us to use the predetermined act and scene structure for linking a quotation from the scholarly work to its source in the literary work. Lastly, we limit the task to the identification of quotations which re-narrate part of the drama either as a summary or a paraphrase.[2] Another unrelated challenge is the acquisition of scholarly works. As opposed to classical dramas, many scholarly works are not readily available online and need to be manually collected, digitized, and cleaned in a very time-intensive process which we outline in Section 4.1.

Our contributions are twofold. Firstly, we present a two-step approach for the identification of indirect quotations, more precisely, summaries and paraphrases, in scholarly works and the source of the quotation in the literary work.[3] In the first step, we identify sentences in the scholarly work that are candidates for containing an indirect quotation. In the second step, we identify the scene of the associated drama which is most likely the source of the quotation. To acquire training data without manual annotation, we use GPT-generated (OpenAI, 2023) summaries as a basis to generate training data for candidate identification and scene prediction. This two-step approach is necessary due the nature of how we generate the training data without manual annotation. Our second contribution is a first dataset of annotated scholarly works with annotations of text spans which summarize or paraphrase the interpreted drama and the source of the quotation.[4]

The paper is organized as follows: The next section gives an overview on related work. In Section 3, we present our method followed by a description of our data acquisition process, the experiments, and results in Sections 4, 5, and 6, respectively. We conclude this work with a discussion in Section 7.

2 Related Work

The task of identifying speech, thought, and writing in fiction and non-fiction texts, referred to as *quotation detection*, is related to the first step of our approach, that is, the identification of summaries or paraphrases in scholarly works. There are different types of speech, thought, and writing, for example, *direct*, *indirect*, or *reported speech* (Semino and Short, 2004; Brunner, 2015). The last type is closest to the scholarly citations in our texts. Quotation detection is often focused on English newspaper articles (Pareti et al. (2013); Scheible et al. (2016)), though there is a corpus-agnostic approach (Papay and Padó, 2019) and an annotated dataset of Finish news articles (Janicki et al., 2023). Corpora for German include (Krug et al., 2018; Brunner et al., 2020a; Petersen-Frey and Biemann, 2024). As part of the Redewiedergabe project,[5] Brunner et al. (2020b) published a number of models for tagging different types of speech in German texts, including one for *reported speech*. A related task is *quotation attribution*, that is, identifying the source of a quotation, for instance, the speaker (Elson and McKeown, 2010; Almeida et al., 2014; He et al., 2013; Muzny et al., 2017).

Although our phenomenon of interest is similar, it is still not easily transferable. Scholarly texts can be quite different in style compared to fictional works or newspaper articles.

The second part of our task is to link quotations to their source. Multiple efforts have been made to understand how attention values of transformer models could be used to identify the source of a summary. Bibal et al. (2022) give an extensive overview on the ongoing debate whether or not attention values can be used to explain black box transformer models. For abstractive summarization specifically, Baan et al. (2019) find that attention values cannot be reliably used to explain summaries. One explanation for these findings could be shortcut learning (Du et al., 2023). Suhara and Alikaniotis (2024) present an approach based on perplexity gain to identify the source of a quotation. They found this method to outperform the second best approach, similarity-based methods, on the XSum dataset (Narayan et al., 2018), while similarity-based methods perform better on the CNN/Daily Mail dataset (Hermann et al., 2015).

Given that our texts are quite different, these results cannot easily be applied to our task. Due to its versatility and availability through SentenceTransformers (Reimers and Gurevych, 2019), semantic textual similarity emerged to be the most promising path. Although there are models which outperform SentenceTranformers (Peng et al., 2022), we decided to use a pre-trained SentenceTransformer (PST) due to the need for German models, which are readily available, and the relative ease of further training due to good documentation and support of a multitude of different use cases.

3 Methods

We first define the task, then describe our approach for generating training data and the training procedure, and then present our tool for inference.

3.1 Task

Our goal is to identify indirect quotations, more precisely, summaries and paraphrases, in scholarly works and link those to the act and scene of the drama which contain the source of the quotation. We divide this into two steps: candidate identification and scene prediction. In the first step, the scholarly work is split into sentences and each sentence is classified as a candidate for (not) containing an indirect quotation. In the second step, for each candidate the most likely source scene in the drama the scholarly work is interpreting is predicted.

[2] For the sake of brevity, we use *quotation* to refer to indirect quotations in the form of summaries and paraphrases.

[3] The source code is licensed under the Apache License 2.0 and available at `https://hu.berlin/indiquo`.

[4] The data is available at `https://doi.org/10.5281/zenodo.15013794` with restricted access due to copyright law.

[5] `http://www.redewiedergabe.de/`

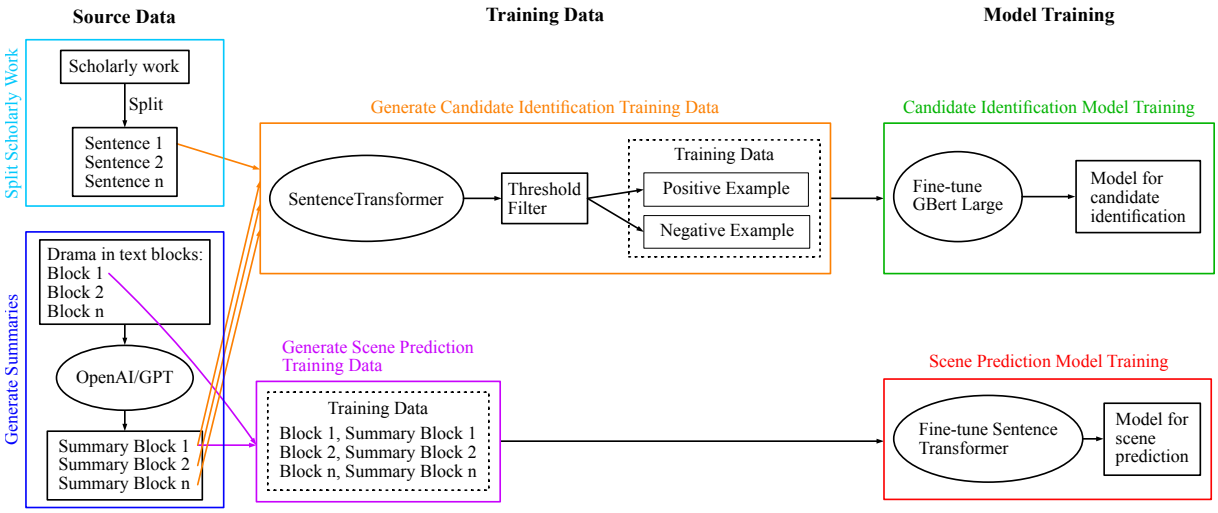

Figure 1: Method Overview

3.2 Training Data Generation

Figure 1 gives an overview of our data collection and training procedure. We assume the dramas to be available as TEI/XML (TEI Consortium, eds., 2022) files in DraCor format and the scholarly works in plain text. The general idea is to use scholarly works, split into sentences (**light blue box**), and drama summaries generated by GPT (**blue box**) as a starting point to generate training data for two models, one for binary classification for candidate identification (**green box**) and one for scene prediction (**red box**). The summarization generation is described in detail in Section 3.2.1.

The dataset for candidate identification contains sentences from scholarly works which are classified as positive, negative, or unclassified based on their similarity to any summary determined by a pre-trained SentenceTransformer (PST) for paraphrase identification (**orange box**). The resulting dataset is used to fine-tune a German BERT model (Devlin et al., 2019) for binary classification (**green box**).

We found that using a PST directly does not clearly outperform a binary classification model on filtered examples (cf. Section 6.3) and has the disadvantage that summaries for every drama are required. We also explored whether summaries could be used directly to fine-tune a PST to improve candidate identification, but found that this would only degrade performance (details in Appendix F).

The dataset for scene prediction consists of pairs of blocks of text from a drama and corresponding summaries (**pink box**). This data is then used to fine-tune a PST for scene prediction (**red box**).

In Sections 3.2.2 and 3.2.3 we describe the training data generation. The resulting models for candidate identification and scene prediction are used at inference time as described in Section 3.4.

3.2.1 Summary Generation

We use the OpenAI API and gpt-4-1106-preview with the following system prompt to generate summaries:[6]

> You are a system for summarizing drama texts. You receive a text and create a short summary of 2-3 sentences. [1]

We left other parameters at their defaults: *temperature* of 1, *top_p* of 1 and *frequency* and *presence penalty* of 0. The maximum number of returned tokens is limited to 200 which should be enough for 2-3 sentences. The user prompt is the text block from the drama without any additional text.

The drama is processed scene by scene. For each scene, speaker turns are concatenated to create text blocks of a maximum length of 128 tokens. For single turns, which are longer than the maximum length, multiple blocks of up to 128 tokens are created. If the last block is shorter than 10 tokens, it is discarded. Stage directions in and between dialogue are included but not scene descriptions. We discuss these decisions in Appendix E.

3.2.2 Candidate Identification Datasets

To generate training data for the candidate identification models, the scholarly texts are split into sentences using Pysbd (Sadvilkar and Neumann, 2020) after footnotes are removed.[7] The sentences are then further processed to make sure that text blocks have a length between 10 and 64 tokens, if possible.[8] This is done by concatenating neighboring sentences until the minimum length is reached without going over the maximum length. If a single sentence is longer than the maximum length, it

[6] Prompt translated from German. All translated texts are followed by a number in brackets which identifies the original text in Appendix A.

[7] We here always only use the running text because footnotes add noise and pose their own challenges.

[8] We use white space tokenization.

is split into parts of the maximum length.[9] With this approach, there are cases where we can end up with sentences which are shorter than the minimum length. As there is no simple solution, we allow such cases for this work. This procedure is necessary as our texts are digitized using OCR with only little manual cleaning. Without merging we end up with too many short partial sentences due to OCR errors or parentheses. Every sentence is then compared to every GPT summary using a German PST for paraphrase identification[10] and the examples are determined as follows:

$$\begin{cases} \text{positive example} & \text{if } \max_{s,t} \text{sim}(s,t) > 0.7 \\ \text{negative example} & \text{if } \max_{s,t} \text{sim}(s,t) < 0.3 \\ \text{unclassified} & \text{otherwise} \end{cases}$$

where $\text{sim}(s,t)$ represents the cosine similarity score between summary s and text block t from the scholarly work. The thresholds were determined using our validation texts (see Section 5.1).

From this data, we create four training datasets. The first contains the positive and negative examples without any modification. The second and third contain examples embedded into their context from the scholarly work. The maximum length of an example is limited to 128 and 256 tokens, respectively. For a fourth, we extend our second dataset with a subset of the data from the Redewiedergabe corpus (Brunner et al., 2020a). We use all instances of type *reported* from texts of type *report* or *review*. This is done to test whether data, that is somewhat similar to our training instances, could help improve the model without additional annotation.

All datasets are balanced between positive and negative instances. Using all available instances would result in an imbalance of about one positive to five negative instances. To get balanced datasets, we randomly downsample the negative examples. Testing different ratios of imbalance did not bring clear improvements to the results.

3.2.3 Scene Prediction Datasets

The foundation for the training data for the scene prediction model is the data collected in Section 3.2.1, that is, pairs of text blocks from the drama and the corresponding GPT summaries. From this data, we create three training datasets.

The first dataset is the collected data without any modification, that is, drama excerpts and the corresponding GPT summaries. For the second dataset, either the original data, that is, drama excerpt and summary, is used as the example, or the GPT summary is split into sentences and the drama excerpt is paired with individual sentences of the summary in order to simulate shorter summaries. For summaries with two sentences, each sentence is paired with the drama excerpt, resulting in two training examples. For summaries of three sentences, also two training examples are created. Either the first two or the last two sentences are concatenated and paired with the drama excerpt, and the remaining sentence is used for the second example. The decision whether to split the summary and which combination to use in the case of a three sentence summary, is made randomly.

The third dataset is like the second but the summary is embedded into random text from scholarly works to see if this makes the model more robust to noise and the specific style of scholarly texts.

3.3 Training

3.3.1 Candidate Identification Model

For each dataset, we fine-tune a German BERT large model[11] with a linear layer on top of the pooled output for binary classification.

3.3.2 Scene Prediction Model

For each dataset, we fine-tune a PST for paraphrase identification with multiple negatives ranking loss (Henderson et al., 2017) to learn the similarity between drama excerpts and summaries.

3.4 Inference

The drama is input as a DraCor XML file and the scholarly work as a plain text file.

3.4.1 Step 1: Candidate Identification

We split the scholarly text into sentences and use the candidate identification model to identify sentences which are quotations.

3.4.2 Step 2: Scene Identification

Using the scene prediction model, we compare every sentence which was classified as a quotation against all text blocks from the drama to identify the most likely origin. We return the act and scene of the text block with the highest similarity as the source.

4 Data

4.1 Acquisition and Digitization

We selected the top 11 dramas with the highest numbers of scholarly interpretations in the online version of the *Bibliographie der deutschen Sprach- und Literaturwissenschaft* (BDLS).[12] This database has a focus on German philology and lists works published since 1985. We excluded *Faust* and *Die Räuber*[13] from the top 11 and collected all scholarly interpretations since

[9] For simplicity, we will still refer to blocks of text as *sentences*. Also, for the remainder of this work, *sentence splitting* always refers to this approach.

[10] https://huggingface.co/deutsche-telekom/gbert-large-paraphrase-cosine

[11] https://huggingface.co/deepset/gbert-large

[12] https://www.bdsl-online.de/

[13] *Faust* was excluded as it has more than six times the number of scholarly interpretations than the next most interpreted drama, *Dantons Tod*. *Die Räuber* was excluded due to an encoding issue with the umlaut during the PDF acquisition.

	Annotators	Precision	Recall	F_1-score
Dantons Tod	A_1/A_2	.52	.25	.33
	A_1/A_3	.67	.24	.34
	A_2/A_3	.72	.47	.56
Iphigenie auf Tauris	A_1/A_2	.50	.34	.38
	A_1/A_3	.63	.41	.46
	A_2/A_3	.59	.47	.51

Table 1: The inter-annotator agreement of the span annotations, measured at the sentence level.

	Annotators	Precision	Recall	F_1-score
Dantons Tod	A_1/A_2	.73	.73	.73
	A_1/A_3	.69	.70	.69
	A_2/A_3	1.00	.97	.98
Iphigenie auf Tauris	A_1/A_2	.72	.69	.70
	A_1/A_3	.98	.78	.83
	A_2/A_3	.91	.90	.90

Table 2: Agreement of scene annotations between annotators at the scene level.

1985 up until the date of collection in 2020 for the remaining dramas. For more details on the corpus, see Appendix B.

All entries from BDSL are manually checked and the PDF downloaded, if available online. The files are then converted to DOCX using Abby FineReader 15. Title pages, headers, and footers are removed; footnotes are not reliably detected and have to be manually checked. The DOCX files are then converted to TEI/XML.

4.2 Annotation

From the 11 dramas, we selected *Dantons Tod* and *Iphigenie auf Tauris* for annotation and to evaluate our experiments. This decision was based on the fact that they differ from each other in their dramatic form. Goethe's *Iphigenie* is a classical, antique-like drama with blank verse while *Dantons Tod* is written in prose without verse. Three people with a background in literary studies annotated the same five scholarly texts for each drama. In addition, another ten texts, five for *Dantons Tod* and five for *Iphigenie auf Tauris*, were each annotated by one annotator. The texts were selected randomly to cover a range of years of publication.

4.2.1 Procedure

The annotation process consisted of two steps. In the first step, the annotators were asked to tag spans of text which are summaries or paraphrases of the literary work. The annotations were done in TEI/XML files without any limitation on the extent of the tagged span. In a second step, the source of the just annotated text spans, that is, the underlying literary text that is summarized or paraphrased, was annotated. This was done by giving line or paragraph numbers, either as single numbers or as ranges. Multiple ranges were allowed.[14]

4.2.2 Results

Overall, the number of annotated instances varies a lot between texts and annotators, from 2 to 61 instances. The numbers also show that two persons systematically annotated more than the third. For more details on the annotations, see Appendix C.

The F_1-score inter-annotator agreement for the span annotation task is shown in Table 1, along with precision and recall. Agreement is calculated on the sentence level and between all combinations of two annotators. To map span annotations to sentences, we take all sentences as positives example which overlap with at least one annotated span and all other sentences as negatives example. Precision is calculated as the ratio of sentences annotated by the first annotator that were also annotated by the second annotator. Recall is the ratio of sentences annotated by the second annotator that were also annotated by the first annotator. On average, annotator 2 and 3 have the highest agreement but it is still relatively low. It should be noted, that the agreement varies a lot between scholarly texts. For some texts, the annotations from one annotator are almost a complete subset of the annotations from the other annotator. Other times, the annotations overlap but both annotators also annotated instances which the other did not. Some of the difficulties are discussed below.

Table 2 shows the F_1-score inter-annotator agreement of the second annotation step. The agreement is calculated on the subset of all annotated spans which overlap with at least one span from the other annotator's annotations. The agreement is calculated on the scene level. Precision and recall are calculated as the ratio of scenes listed by both annotators to the number of scenes listed by the first and second annotator, respectively.

Overall, the agreement for this second step is a lot higher. Again, the agreement for the second and third annotator is highest on average. The agreement also varies between texts but is overall more stable.

From the individual annotations, a gold standard was created in consultation between the three annotators. During this process, reasons for the discrepancies were discovered that we describe in the next section.

4.2.3 Challenges

For the first step of the annotation task, a first challenge arises from the fact that interpretive texts often do not clearly distinguish between quotations and other references to the literary text. Generally, text passages that contain exegesis, interpretation, and other forms of explanation of the literary text contain some form of reference simply because the literary critic necessarily has to refer to the literary text to interpret it. Consequently, one of the challenges was to identify "pure" indirect quotations, that is, summarizations and paraphrases, without interpretive parts that stem from the author. The following example illustrates such a case:

[14] Annotation Guidelines: https://doi.org/10.5281/zenodo.15006101

> The happy resolution of the conflict on Tauris is only possible here through the disclosure of all plans, i. e., through the courage to tell the truth. [2]

On the surface, the whole sentence could be seen as an indirect quotation but looking at the indiviual parts, we can observe that only the phrase "the disclosure of all plans" should be considered an indirect quotation of the action of one of the dramatis personae. Whereas "The happy [...] conflict" refers to the drama as a whole and the phrase "i. e., through [...] truth" is the critic's interpretation. A similar problem arises in passages where direct and indirect quotations are merged in one sentence:

> While he orders her to carry out her service at the end of the first encounter (I,3:537), she presents him with the imperative refusal of command in V,3: 'Spoil us, if you may'! (Vs. 1936), which leads to an instruction at the end of the scene: 'Consider not; grant as you feel' (Vs. 1992). [3]

In the middle part of the passage, we can find both, an indirect ("she presents him [...]") and a direct quotation of the drama. There is a double reference because the indirect quotation announces the following direct quotation, which can be considered a quotation of a quotation. The examples illustrate how the nature of interpretive texts makes the identification of indirect quotations very hard and often leads to ambiguous cases which are difficult and time-consuming to classify, even for human experts.

In the second step, one challenge is to identify how narrow or wide the source annotation should be. Usually, the exact extent of the annotation will not change the scene and therefore the agreement between annotators is not affected. A second challenge arises from indirect quotations which do not refer to a specific part of the drama but are broader and sometimes even reference the whole drama. These cases are also difficult with regard to the first annotation step as we are not interested in quotations which are too broad. A third challenge stems from the fact that interpretive texts can be inaccurate in recapitulating passages of the drama:

> The conflict inside her escalates into agony when she recognizes her brother in one of the strangers to be sacrificed. [4]

The sentence refers to two very different parts of the play which is not easy to figure out. As a result, the sentence had to be annotated with two different verse sources. Merging information from disparate parts of the drama in one indirect quotation is a practice the annotators observed more than once.

5 Experiments

5.1 Scholarly Texts Split

The 20 texts are partitioned into four sets. The first set (*Dev*) contains six randomly selected texts for preliminary experiments and to determine thresholds. The second set (*Gold*) contains the remaining texts from our gold annotations which were not used for validation. The third set (*Single*) contains the texts which were only annotated by one annotator. Finally, the fourth set (*Few*) contains texts with five or fewer instances.

5.2 Training and Validation Datasets

All datasets are created from dramas and corresponding scholarly texts which are not used for testing, that is, *Dantons Tod* and *Iphigenie auf Tauris* are not used in any of the training and validation datasets. We split all datasets into 90 % training and 10 % validation instances.

5.3 Training and Evaluation Metrics

We evaluate on the sentence level. Every sentence that has any overlap with an annotated span in our gold corpus is a positive example.

5.3.1 Candidate Identification

We compare four variants of the candidate identification model against two baseline models, a pre-trained SentenceTransformer (*Baseline-ST*) and the tagger for reported speech from the Redewiedergabe project (*Baseline-RW*). The four variants are each trained on one of the datasets described in Section 3.2.2: The examples without additional context (*No-Context*), the examples with context, limited to 128 tokens (*Context-128*) and 256 tokens (*Context-256*), and with additional examples from the Redewiedergabe corpus (*Context-128-RW*).

For the first baseline, the scholarly work is split into sentences. Every sentence is then compared to all text blocks from the drama and a sentence is classified as a summary if at least one drama/summary pair is above a threshold of 0.5. For the second baseline, we map the results from the Redewiedergabe tagger to the sentences from the scholarly work by classifying a sentence as a summary if any part of that sentence was tagged as reported speech by the tagger.

Each variant of our model was fine-tuned for five epochs with a batch size of 16 and a learning rate of $2 \cdot 10^{-5}$. We use a classification threshold of 0.5 for all model variants. During pretests using the validation scholarly works, we found the ideal threshold to vary a lot depending on the scholarly work and 0.5 was the only reasonable choice based on the small number of texts. For the evaluation we use the checkpoint with the best F_1-score on the validation split of the dataset.

5.3.2 Scene Prediction

We compare three variants of the scene prediction model against a pre-trained SentenceTransformer (*Base*) as the baseline. The three variants are each trained on one of the datasets described in Section 3.2.3: The drama excerpts with summary (*Long*), the drama excerpts with short summaries (*Short*), and the drama excerpts with short summaries embedded into text (*Short-Emb*).

Each variant of our models was fine-tuned for five epochs with a batch size of 16 and a learning of $2 \cdot 10^{-5}$. For the evaluation we use the checkpoint with the best average precision on the validation split of the dataset.

6 Results

6.1 Candidate Identification

Results are shown in Table 3. *Context-128* performs best on the *Dev*, *Gold*, and *Single* set with an F_1-score of 0.37, 0.31, and 0.39, respectively. The baselines are outperformed on all sets except *Gold*, where only *Context-128* performs better. Texts with five or less instances (*Few*) have the worst results due to very low precision (though recall is on the same level as for the other sets). Precision is relatively low overall.

As we have seen, the performance depends on the number of instances in the scholarly text and the set of scholarly works. To understand whether the nature of the sets or individual works are the reason, we report the F_1-score for the five *Gold* texts in Table 4. *Baseline-ST* outperforms our approach on two texts and on *Pet06*[15] both baselines outperform our approach. On *Hoe06*, the performance is close to the baselines. *Context-128* outperforms both baselines on three texts. The variance in performance is less pronounced for the baselines. We conclude that the performance heavily depends on the individual scholarly work and, to a lesser extent, also on the model. We observed similar effects for other scholarly works during development.

In conclusion, looking at the results in isolation they do not seem very promising. Comparing the results to the inter-annotator agreement, we get a better idea of their relative quality: the highest agreement we get is 0.56 for annotators 2 and 3 for *Dantons Tod* and 0.51 for *Iphigenie auf Tauris*.

Error analysis One source of the low precision could be the way in which the training data is generated and that this process leads to data that contains too many false positives. We described the process in Section 3.2.2 with a lower and upper threshold of 0.3 and 0.7, respectively. These result in 122 true negative examples on our development set and no false negatives. But the upper threshold of 0.7 generates 46 true positives and 79 false positives. Upon manual analysis we found among them many edge cases, similar to the difficult cases identified during annotation, and using a higher threshold would lead to too few examples overall.

We also identified some issues related to specific characteristics of the scholarly works. *Bor09*, for example, compares, and therefore references, a number of different adaptations of Iphigenie (Schiller, Euripides (taurische Iphigenie), Gluck's Iphigenie). This results in a lot of passages which renarrate the story of Iphigenie but do not quote Goethe's Iphigenie and this in turn results in a high number of false positives. The scholarly works often reference more dramas than just the one which is the main focus of the interpretation. This is, for example, the case with *Pet06* and *Cam19*. This again, results in a high number of false positives.

6.2 Scene Identification

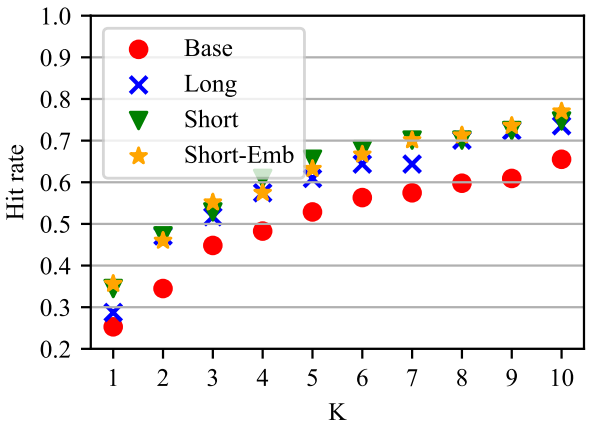

Figure 2: Scene evaluation on the *Gold* set.

Figure 2 reports hit rate at K for the top 1 to 10 scenes. All three variants outperform the baseline, with the general trend that *Short* and *Short-Emb* achieve a higher performance than *Long*.

To evaluate how the performance varies between the sets of scholarly works, we compare the performance of the *Short* model in Figure 3. As before, we notice a varying performance between sets which is lowest for *Gold* and highest for *Dev*.

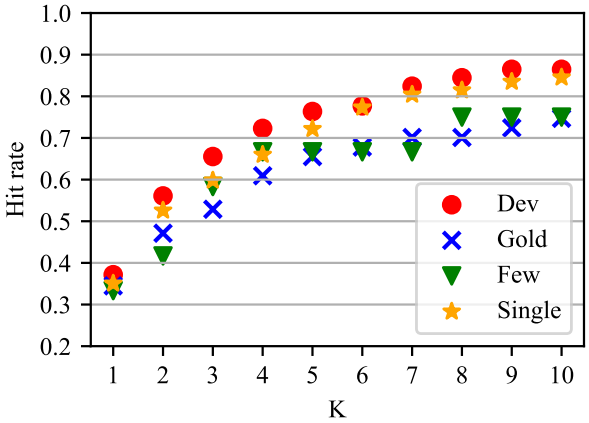

Figure 3: Scene evaluation of the different sets.

Again, to confirm that the underlying reason for this is not the nature of the sets but individual scholarly works, we compare the performance of the individual texts of the *Gold* set in Table 5. Again, performance varies between the texts and *Pet06* has the lowest for the baseline and all model variants. The model *Short-Emb* outperforms the baseline on all texts for HR@10 but for some texts the other two variants perform better, for

[15] Texts are labeled with the first (up to three) letters of the first author's name followed by the last two digits of their year of publication. The labels can be used to identify the texts on https://hu.berlin/quidex-en.

Approach	Dev	Gold	Single	Few
Baseline-ST	.19/.49/.28	.20/.63/.30	.20/.59/.29	.03/.33/.05
Baseline-RW	.16/.41/.23	.11/.26/.16	.15/.34/.21	.03/.33/.05
No-Context	.26/.54/.35	.23/.30/.26	.29/.51/.37	.04/.33/**.08**
Context-128	.25/.68/**.37**	.23/.45/**.31**	.28/.65/**.39**	.05/.58/**.08**
Context-128-RW	.25/.62/.36	.23/.41/.30	.25/.60/.35	.05/.50/**.08**
Context-256	.25/.68/**.37**	.23/.40/.29	.26/.62/.37	.05/.50/**.08**

Table 3: Precision, recall, and F_1-score for candidate identification.

Approach	Pet06	Kos08	Mro13	Hoe16	Bur17
Baseline-ST	.40	.25	.20	.40	.29
Baseline-RW	.21	.22	.19	.08	.06
No-Context	.00	.34	.00	.38	.16
Context-128	.06	.33	.25	.32	.48
Context-128-RW	.00	.32	.17	.35	.44
Context-256	.07	.31	.00	.36	.44

Table 4: F_1-score for texts of the *Gold* set.

example, *Long* performs best for *Pet06* for HR@5 and HR@10.

Approach	Pet06	Kos08	Mro13	Hoe16	Bur17
Base	.27/.33	.64/.79	.60/.73	.58/.79	.57/.67
Long	.33/.61	.43/.64	.80/.80	.79/.89	.67/.71
Short	.33/.44	.71/.86	.80/.80	.68/.74	.76/.90
Short-Emb	.22/.50	.64/.86	.73/.87	.79/.84	.76/.81

Table 5: Hit rate (HR@5/HR@10) for the *Gold* set.

6.3 Ablation

To generate training data for candidate identification, we use a PST and GPT-generated summaries to identify positive and negative examples. We use an upper and a lower threshold to find examples where the model assigns relatively low and high scores, respectively. This raises the question if it would be possible to use this approach to identify candidate sentences directly, that is, replace the lower and upper thresholds with a single threshold, and compare sentences with GPT summaries. Additionally, we can also use the score returned by the PST for the scene prediction step. This is the same as our normal scene prediction step but instead of comparing sentences with drama excerpts, we compare sentences to summaries of drama excerpts.

For the candidate identification, we determine the best threshold of 0.655 on the development set and get the following F_1-scores: Dev/Gold/Single/Few: 0.38/0.36/0.35/0.13. The results are overall more stable over the different sets of scholarly works but our approach is not clearly outperformed.

For scene prediction, the results are reported in Figure 4. The performance is better than our approach across all datasets. A reason for this could be that summaries are closer to the types of text the PST was trained on than drama excerpts.

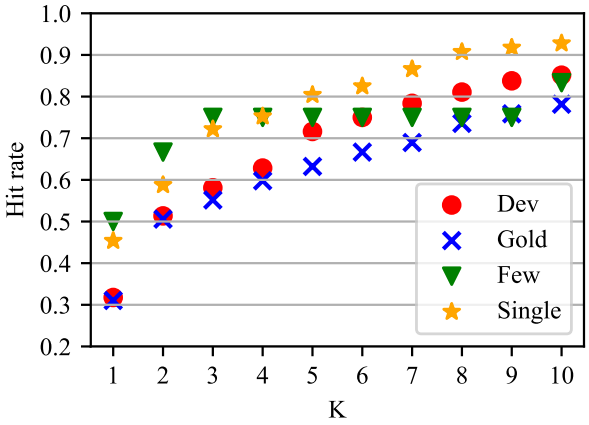

Figure 4: PST summary scene evaluation.

7 Conclusion

Our contributions are twofold. First, we created a dataset of scholarly works with annotations of text spans which summarize or paraphrase a drama, and their source in the drama. We created a gold standard from three independent annotations. During this process, we identified several reasons for discrepancies in the annotations and the resulting inter-annotator agreement. Second, we presented an approach for the automatic identification and linking of indirect quotations in scholarly works and dramas leveraging semantic similarity. We chose the approach as we hoped the trained model would allow us to work with arbitrary dramas without the need for summaries at inference time. We evaluated the approach and identified different challenges.

For candidate identification we found distinguishing between indirect quotations and surrounding text very difficult, even for human experts. One reason is that scholarly texts necessarily reference the literary work, and that these references can take various shapes and forms which are hard to separate, especially because interpreting passages can be quite similar to summaries and paraphrases. We also found the performance of the model to heavily depend on the specific text. Two reasons for this are that some texts discuss multiple adaptions of the same work and that some discuss sev-

eral dramas.

In light of these results, we evaluated whether we could perform better if we assume summaries to be available. We found that another approach which uses summaries instead of drama excerpts performs more stable overall, and in the case of scene prediction, also outperforms our approach. A likely reason is that summaries are closer in style to the types of text language models are normally trained on.

We conclude that the main area for improvement is the identification of semantic similarity in the context of indirect quotations, which existing models can not fully capture due to the similarity between relevant references and the surrounding text. At least in part, this might be due to named entities. Hatzel and Biemann (2024b,a) find that models for semantic similarity strongly rely on named entities as a source of similarity. Consequently, information on the argumentation structure of the scholarly work is needed for a better distinction. Finally, the limited input size of models such as BERT, which necessitates splitting of texts, is another challenge and area for future work.

With the increasing performance of recent large language models (LLM) on a variety of tasks and the increasing context window size, another route for the future will be to utilize LLMs in a more direct fashion and prompt with full scholarly and literary texts to extract indirect quotations.

8 Limitations

Our dataset has different limitations. Firstly, all dramas are written by male authors. We are limited with regard to the dramas we can use for our experiment by the availability of scholarly works for these dramas. Secondly, our annotated dataset is quite small with 20 annotated scholarly works of which half were annotated by multiple annotators. Additionally, our dataset has limited variety as we only annotated scholarly works from two dramas. Our approach is also limited to literary texts for which a suitably granular segmentation is available, for example, the act and scene structure of dramas. In addition, our further segmentation of the literary and scholarly texts is not ideal and can be improved, see Appendix E for more details.

Automatic generation of summaries using GPT introduces limitations. For example, we found that stylistic differences between GPT summaries and scholarly works introduce issues when fine-tuning a PST, see Appendix F.

Lastly, we assume scholarly works to be available in digitized form as plain text. Transforming PDF files into this form is a time and resource intense process and involves a number of manual steps in case the quality of the PDF files is low.

Acknowledgments

Parts of this research were funded by the German Research Foundation (DFG) priority programme (SPP) 2207 *Computational Literary Studies* project *Is Expert Knowledge Key? Scholarly Interpretations as Resource for the Analysis of Literary Texts in Computational Literary Studies* (grant no. 424207720). We would like to thank the project's student assistants Marielena Rasch, Elisabeth Renger and Gregor Sanzenbacher for their excellent work despite the difficulty of the annotation tasks.

References

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48, Gothenburg, Sweden. Association for Computational Linguistics.

Frederik Arnold and Robert Jäschke. 2021. Lotte and Annette: A Framework for Finding and Exploring Key Passages in Literary Works. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 55–63, NIT Silchar, India. NLP Association of India (NLPAI).

Frederik Arnold and Robert Jäschke. 2023. A novel approach for identification and linking of short quotations in scholarly texts and literary works. *Journal of Computational Literary Studies*, 2.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization? In *Proceedings of the SIGIR Workshop FACTS-IR*.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Annelen Brunner. 2015. *Automatische Erkennung von Redewiedergabe. Ein Beitrag zur quantitativen Narratologie*. De Gruyter, Berlin, München, Boston.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020a. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 803–812, Marseille, France. European Language Resources Association.

Annelen Brunner, Ngoc Duyen Tanja Tu, Lukas Weimer, and Fotis Jannidis. 2020b. To bert or not to bert-comparing contextual embeddings in a deep learning architecture for the automatic recognition of four types of speech, thought and writing representation. In *SwissText/KONVENS*.

Robert-Alain de Beaugrande and Wolfgang Ulrich Dressler. 1981. *Einführung in die Textlinguistik*. Max Niemeyer Verlag, Tübingen.

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 of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Commun. ACM*, 67(1):110–120.

David K. Elson and Kathleen R. McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, page 1013–1019. AAAI Press.

Frank Fischer, Ingo Börner, Mathias Göbel, Angelika Hechtl, Christopher Kittel, Carsten Milling, and Peer Trilcke. 2019. Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities", Utrecht, July 9–12, 2019*. Utrecht University.

Hans Ole Hatzel and Chris Biemann. 2024a. Story embeddings — narrative-focused representations of fictional stories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5931–5943, Miami, Florida, USA. Association for Computational Linguistics.

Hans Ole Hatzel and Chris Biemann. 2024b. Tell me again! a large-scale dataset of multiple summaries for the same story. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15732–15741, Torino, Italia. ELRA and ICCL.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Rudolf Helmstetter. 2003. Zitat. In Jan-Dirk Müller, editor, *Reallexikon der deutschen Literaturwissenschaft*, volume 3, pages 896–899. De Gruyter, Berlin, New York.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Maciej Janicki, Antti Kanner, and Eetu Mäkelä. 2023. Detection and attribution of quotes in Finnish news media: BERT vs. rule-based approach. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 52–59, Tórshavn, Faroe Islands. University of Tartu Library.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus, and Fotis Jannidis. 2018. Description of a Corpus of Character References in German Novels - DROC [Deutsches ROman Corpus]. In *DARIAH-DE Working Papers*. DARIAH-DE.

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, Valencia, Spain. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sean Papay and Sebastian Padó. 2019. Quotation detection and classification with a corpus-agnostic model. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 888–894, Varna, Bulgaria. INCOMA Ltd.

Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 989–999, Seattle, Washington, USA. Association for Computational Linguistics.

Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. 2022. Predicate-argument based bi-encoder for paraphrase identification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland. Association for Computational Linguistics.

Fynn Petersen-Frey and Chris Biemann. 2024. Dataset of quotation attribution in German news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4412–4422, Torino, Italia. ELRA and ICCL.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745, Berlin, Germany. Association for Computational Linguistics.

Elena Semino and Mick Short. 2004. *Corpus Stylistics: Speech, Writing And Thought Presentation In A Corpus Of English Writing*. Routledge, London/New York.

Yoshi Suhara and Dimitris Alikaniotis. 2024. Source identification in abstractive summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–224, St. Julian's, Malta. Association for Computational Linguistics.

TEI Consortium, eds. TEI P5: Guidelines for electronic text encoding and interchange, version 4.4.0 [online]. 2022.

Simone Winko. 2022. Bezugnahmen auf die Textwelt. Untersuchungen zu Handlungstypen in der literaturwissenschaftlichen Interpretationspraxis. *Scientia Poetica*, 26(1):125–166.

A Translations

1. You are a system for summarizing drama texts. You receive a text and create a short summary of 2-3 sentences.

   Du bist ein System zur Zusammenfassung von Dramentexten. Du bekommst einen Text und erzeugst eine kurze Zusammenfassung von 2-3 Sätzen.

2. The happy resolution of the conflict on Tauris is only possible here through the disclosure of all plans, i. e. through the courage to tell the truth.

   Die glückliche Lösung des Konfliktes auf Tauris ist hier nur möglich durch die Offenlegung aller Pläne, d. h. durch den Mut zur Wahrheit.

3. While he orders her to carry out her service at the end of the first encounter (I,3:537), she presents him with the imperative refusal of command in V,3: 'Spoil us, if you may'! (Vs. 1936), which leads to an instruction at the end of the scene: 'Consider not; grant as you feel' (Vs. 1992).

   Während er ihr am Ende der ersten Begegnung den Befehl erteilt, ihren Dienst auszuüben (I,3:537), präsentiert sie ihm in V,3 die imperative Befehlsverweigerung: 'Verdirb uns, wenn du darfst'! (Vs. 1936), was am Schluss der Szene in eine Handlungsanweisung mündet: 'Bedenke nicht; gewähre, wie du's fühlst' (Vs. 1992).

4. The conflict inside her escalates into agony when she recognizes her brother in one of the strangers to be sacrificed.

   Der Konflikt in ihrem Inneren steigert sich zur Höllenqual, als sie in einem der zu opfernden Fremden den Bruder erkennt.

B Corpus Details

Table 6 gives an overview on the resulting dataset of dramas and the numbers of scholarly texts.

C Annotation Details

Table 7 reports the number of annotated spans. Texts are labeled with the first (up to three) letters of the first author's name followed by the last two digits of their year of publication. The labels can be used to identify the texts on: `https://hu.berlin/quidex-en`.

D Training Details

For the candidate identification, we experiment with four datasets, as described earlier. For the first three datasets the split results in datasets with 4 648 training instances and 516 validation instances. For the fourth dataset we end up with 6233 training and 691 validation instances. Half of the additional instances are from the Redewiedergabe corpus and the other half are negative examples to balance the dataset.

For the scene prediction experiments, the first dataset contains 1 927 training and 214 validation instances. The other two datasets both contain 3 192 training and 354 validation instances.

E Segmentation Details

For the generation of summaries, we currently do not include scene descriptions. This is not an issue for *Dantons Tod* and *Iphigenie auf Tauris* with very few scene descriptions, but could affect other dramas and it should be further investigated how this influences the results. Lastly, when the drama is split into blocks, the speaker is only part of the first block if a single turn is longer than the maximum length. It could make sense to add the speaker to subsequent blocks.

F Alternative Candidate Identification

For candidate identification, we explored an alternative approach using a dataset of positive examples, which are blocks of text from the drama and the corresponding summary, and negative examples, which are the same

Title	Author	Year	Texts
Dantons Tod	Georg Büchner	1835	76
Emilia Galotti	Gotthold Ephraim Lessing	1772	34
Die Hermannsschlacht	Heinrich von Kleist	1808	34
Iphigenie auf Tauris	Johann Wolfgang Goethe	1787	44
Die Jungfrau von Orleans	Friedrich Schiller	1801	26
Leonce und Lena	Georg Büchner	1836	21
Maria Stuart	Friedrich Schiller	1800	23
Nathan der Weise	Gotthold Ephraim Lessing	1779	41
Penthesilea	Heinrich von Kleist	1808	49
Prinz Friedrich von Homburg	Heinrich von Kleist	1821	39
Wilhelm Tell	Heinrich von Kleist	1804	26

Table 6: Names, authors, and publication years of dramas together with the number of scholarly works we found for each drama.

Scholarly work	A_1	A_2	A_3	Gold
Dantons Tod				
Ded92	-	15	-	-
Hil99†	16	28	30	23
Här02†	-	20	-	-
Pet06	6	13	22	20
Hes07	-	4	-	-
Hol13	-	2	-	-
Mro13	3	8	19	14
Bur17	6	12	24	18
Dub17	-	17	-	-
Cam19†	10	20	17	20
Iphigenie auf Tauris				
Gla91	-	-	11	-
Kli95†	-	-	19	-
Fri01	-	-	3	-
Jes05	2	3	3	4
Kos08	2	7	12	12
Bor09†	9	14	19	14
Hor11†	29	60	61	52
Spa14	-	-	23	-
Hoe16	15	17	22	21
Epp18	-	-	29	-

Table 7: Number of annotated spans of the three annotators A_i. † indicates texts used for validation.

block of text from the drama and a random sentence from a scholarly work, to fine-tune a PST to assign a higher similarity to pairs of drama excerpts and an actual summary compared to drama excerpts paired with other text. We found this approach to perform worse than just the PST without any further training. One reason could be that selecting random sentences from the scholarly work introduces too many false examples where the selected sentence is actually a summary. Another reason could be stylistic differences between the texts, that is, GPT summaries and scholarly works, and shortcut learning effects (Du et al., 2023).

This probably also affects the scene prediction step. Different to the identification task, the model cannot just learn stylistic differences between training instances as all data comes from the same sources. It is still likely that the differences between GPT summaries and real scholarly texts reduce performance but there is no readily available alternative.