








Automatic Topic-Guided Segmentation of Holocaust Survivor Testimonies

Eitan Wagner¹ 
Renana Keydar² 
Amit Pinchevski³ 
Omri Abend¹ 

1. School of Computer Science and Engineering, Hebrew University of Jerusalem , Jerusalem, Israel.
2. Department of Law and Digital Humanities, Hebrew University of Jerusalem , Jerusalem, Israel.
3. Department of Communications, Hebrew University of Jerusalem , Jerusalem, Israel.

Citation

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend (2023). "Automatic Topic-Guided Segmentation of Holocaust Survivor Testimonies". In: *Journal of Computational Literary Studies* 2 (1). [10.48694/jcls.3580](https://doi.org/10.48694/jcls.3580)

Date published 2024-02-10

Date accepted 2024-01-12

Date received 2023-02-17

Keywords

segmentation, NLP, spoken narratives, testimonies, narrative analysis, topic analysis, mutual information

License

CC BY 4.0 

Reviewers

Jan Rybicki, Mats Malm

Note

This paper has passed through the conference track of JCLS. In addition to being peer reviewed, it was presented and discussed at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. In recent decades, efforts have been made to gather and digitize the testimonies of living Holocaust survivors. The challenge we now face is attending to those thousands of human stories, which while safely stored in archives, may nevertheless disappear into oblivion. Despite recent advances in narrative analysis in the fields of Computational Literature (CL) and Natural Language Processing (NLP), existing language model technology still faces challenges in analyzing elaborate narratives and long texts. One such challenge is text segmentation – a long-standing issue in the area of CL and NLP. In our work, we propose a computational method to approach this problem. Our research draws on testimony transcripts from the Shoah Foundation (SF) Holocaust archive for supervised topic classification, which is then used as topic guidance for automatic segmentation.

1. Introduction

Major efforts are being devoted to improving digital access to Holocaust testimonies across the world for safeguarding, cataloging, and disseminating survivors' personal accounts. The imminent passing of the last remaining survivors coincides with the transformation from analog platforms (such as film, video, and television) to digital platforms (big data, online access, social media), which introduces great challenges – and great opportunities – to the future of Holocaust memory. As the phase of survivors' testimony collection reaches its inevitable conclusion, pressing questions emerge: How can we approach and make sense of the enormous quantity of materials collected, which by now exceeds the capacity of human reception? How can we study and analyze the multitude of testimonies in a systematic yet ethical manner, one that respects the integrity of each personal testimony? How can new technology help us cope with the gap between mass atrocity and mass testimony (Keydar 2019)?

Advances in the fields of Computational Literature (CL) and Natural Language Processing (NLP) hold the potential of opening new avenues for the computational analysis of testimony narratives at scale. However, notwithstanding recent developments in the field, existing language modeling technology still faces challenges in analyzing elaborate narratives and long texts in general.

One intuitive approach to dealing with long texts is through an intermediate step of segmentation. Although seemingly simple, the notion of segmentation is not easy to define, as there are many different considerations that may favor placing certain boundary points, rather than others.

In our work, we propose to perform segmentation under the guidance of topics, resulting in the task known as *topical segmentation* (Hearst 1997; Kazantseva and Szpakowicz 2012). For topics, we use a set of predefined topical categories, created by domain experts. The task of topical segmentation is well studied, but previous work has mostly addressed it in the context of structured, well-defined segments, such as segmentation into paragraphs, chapters, or segmenting text that originated from multiple sources. We tackle the task of segmenting running (spoken) narratives, which poses hitherto unaddressed challenges. As a test case, we address Holocaust survivor testimonies, given in English. Other than the importance of studying these testimonies for Holocaust research, we argue that they provide an interesting test case for topical segmentation, due to their unstructured surface level, relative abundance (tens of thousands of such testimonies were collected), and the relatively confined domain that they cover. Our work leverages the annotations from the Shoah Foundation (SF) Holocaust testimony archive for supervised topic classification and uses these topics as guidance for the topical segmentation of the testimonies.

In this contribution, we discuss the importance and challenges of narrative segmentation as the basis for the analysis of more complex narratological phenomena and as a method for representing the narrative flow. We follow Reiter (2015) and his notion of narrative segments as a pragmatic intermediate layer, which is a first step towards the annotation of more complex narratological phenomena. We expand Reiter's work in two main directions: 1) We hypothesize that boundary points between segments correspond to low mutual information between the sentences proceeding and following the boundary. Based on this hypothesis, we develop a computational model of segmentation for unstructured texts that has the prospect of being identifiable automatically and theoretically sound. We propose a simple computational method for automatic segmentation. 2) We focus on topic-based segmentation rather than on event-based segmentation (Gius and Vauth 2022). This allows the use of topic models and classifiers which are easier to obtain compared to event models. We provide a set of expert-created Holocaust-related topics and train a classifier with it. We consider it an important addition to recent efforts to operationalize narrative theory in CL and NLP.

Our work also contributes to current Holocaust research, seeking methods to better access testimonies (Artstein et al. 2016; Fogu et al. 2016). We expect our methods to promote theoretical exploration and analysis of testimonies, enabling better access, research, and understanding of the past.

This work presents the following results:

- We define segments in a theoretically sound manner, building upon information-theory measures.
- We propose a simple and effective algorithm for segmentation, independent of topics.

- We evaluate the model and argue for the necessity of guidance for segmentation, especially in unstructured texts.
- We construct data and models for topic classification. We propose models to infer topics and segments as a combined task.
- We show that giving a reference (“gold”) segmentation leads to better topics, but it seems difficult to design a joint model that gives a segmentation that benefits the topics.
- We discuss future directions to address this difficulty.

We note that the technical and algorithmic part of the paper was adopted with minimal changes from Wagner et al. (2022).

2. Approaches to Narrative Segmentation

2.1 Narrative Analysis

Proper representation of narratives in long texts remains an open problem in computational narratology and NLP (Castricato et al. 2021; Mikhalkova et al. 2020; Piper et al. 2021; Reiter et al. 2022). High-quality representations for long texts seem crucial to the development of document-level text understanding technology, which is currently unsatisfactory (Shaham et al. 2022). One possible avenue for representing long texts is to cast them as a sequence of shorter segments, with inter-relations between them.

This direction has deep conceptual roots. Beginning with Aristotle’s theory of drama, narratological analysis has relied on the identification of and distinction between one event and the next. But what guides the segmentation – what constitutes the divide between events – has remained obscure. While some genres offer structural cues for segmentation, such as diary entries, scenes in a dramatic play, stanzas in a poem, or chapters in a novel, other forms of narratives do not always present clear units or boundaries (Gius and Vauth 2022; Zehe et al. 2021).

From the computational perspective, much work has been done in the direction of probabilistic schema inference, focusing on either event schemas (Chambers 2013; Chambers and Jurafsky 2009; M. Li et al. 2020) or persona schemas (Bamman et al. 2013, 2014).

A common modern approach for modeling narratives is as a sequence of neural states (Rashkin et al. 2020; Wilmot and Keller 2020, 2021). Wilmot and Keller (2020) presented a neural GPT2-based model for suspense in short stories. This work follows an information-theoretic framework, modeling the reader’s suspense by different types of predictability. Wilmot and Keller (2021) present another neural architecture for story modeling. Due to their strong performance in text generation, neural models are also commonly used for story generation, with numerous structural variations (Alhussain and Azmi 2021; Rashkin et al. 2020; Zhai et al. 2019). However, a general drawback of the neural approach is the lack of interpretability, which is specifically crucial in the context of drawing qualitative conclusions from experiments.

A different approach represents and visualizes a narrative as a sequence of interpretable topics. Min and Park (2019) visualized plot progressions in stories in various ways, including the progression of character relations. Antoniak et al. (2019) analyzed birth stories, but used a simplistic, uniform segmentation, conjoined with topic modeling, to visualize the frequent topic paths.

Inspired by this approach, we seek to model the narrative of a text using topic segmentation, dividing long texts into topically coherent segments and labeling them, thus creating a global topical structure in the form of a chain of topics. An NLP task of narrative representation of a given document may benefit from knowing something about the document's high-level structure. Topical segmentation is a lightweight form of such structural analysis: Given a sequence of sentences or paragraphs, split it into a sequence of topical segments, each characterized by a certain degree of topical unity (Kazantseva and Szpakowicz 2014). This is particularly useful for texts with little structure imposed by the author, such as speech transcripts, meeting notes, or, in our case, oral testimonies. Topic segmentation can be useful for the indexing of a large number of testimonies (tens of thousands of testimonies have been collected thus far) and as an intermediate or auxiliary step in tasks such as summarization (Jeff Wu et al. 2021) and event detection (Wang et al. 2021).

Unlike recent supervised segmentation models that focus on structured written text, such as Wikipedia sections (Arnold et al. 2019; Lukasik et al. 2020) or book chapters (Pethe et al. 2020), we address the hitherto mostly unaddressed task of segmenting and labeling unstructured (transcribed) spoken language. For these texts, we do not have large datasets of divided text. Moreover, there may not be any obvious boundaries that can be derived based on local properties. This makes the task more challenging and hampers the possibility of taking a completely supervised approach.

To adapt the model to jointly segment and classify, we incorporate into the model a supervised topic classifier, trained over manually indexed one-minute testimony segments, provided by the USC Shoah Foundation (SF).¹ Inspired by Misra et al. (2011), we also incorporate the topical coherence based on the topic classifier into the segmentation model.

2.2 Topic Classification

The *topic* of a text segment is the subject or theme guiding the segment. Latent Dirichlet Allocation (LDA; Blei et al. 2003) is a popular method to extract latent topics in an unsupervised fashion. In LDA, the definition of a topic is a distribution over a given vocabulary. This definition is very flexible, with one of the results being computationally heavy at inference time.

In recent years, attempts have been made, in various degrees of success, to apply topic models for the purpose of narrative analysis prose (Jockers and Mimno 2013; Uglanova and Gius 2020) and drama (Schöch 2017). Topic models can also be useful in the analysis of complex narratives, such as oral testimonies and other free-form narrative texts (Keydar et al. 2022). Despite its popularity, we found LDA to be too heavy computationally for inference on texts with many segments. Therefore we did not apply LDA

1. See: <https://sfi.usc.edu/>.

in our research. Instead, we used supervised text classification over domain-specific predefined topics.

In supervised classification, we have a list of predetermined topics and a set of texts, each assigned a topic from the list. A classifier is trained to predict the topic for a given text. Since the introduction of BERT (Devlin et al. 2018), the common practice in NLP is to use a neural model that was pretrained on general language tasks and finetune it for the downstream task of classification. This method achieves impressive results even without a vast amount of labeled data, thus proving a natural choice for many domains.

2.3 Text Segmentation

Considerable previous work addressed the task of text segmentation, using both supervised and unsupervised approaches. Proposed methods for unsupervised text segmentation can be divided into linear segmentation algorithms and dynamic graph-based segmentation algorithms.

Linear segmentation, i.e., segmentation that is performed on the fly, dates back to the TextTiling algorithm (Hearst 1997), which detects boundaries using window-based vocabulary changes. Recently, He et al. (2020) proposed an improvement to the algorithm, which, unlike TextTiling, uses the vocabulary of the entire dataset and not only of the currently considered segment. TopicTiling (Riedl and Biemann 2012) uses a similar approach, using LDA-based topical coherence instead of vocabulary only. This method produces topics as well as segments. Another linear model, BATS (Q. Wu et al. 2020), uses combined spectral and agglomerative clustering for topics and segments.

In contrast to the linear approach, several models follow a Bayesian sequence modeling approach, using dynamic programming for inference. This approach allows making a global prediction of the segmentation, at the expense of higher complexity. Implementation details vary, and include using pretrained LDA models (Misra et al. 2011), online topic estimation (Eisenstein and Barzilay 2008; Mota et al. 2019), shared topics (Jeong and Titov 2010), ordering-based topics (Du et al. 2015), and context-aware LDA (W. Li et al. 2020).

Following recent advances in neural models, these models have been used for the task of supervised text segmentation. Pethe et al. (2020) presented ChapterCaptor which relies on two methods. The first method performs chapter break prediction based on Next Sentence Prediction (NSP) scores. The second method uses dynamic programming to regularize the segment lengths toward the average. The models use supervision for finetuning the model for boundary scores, but can also be used in a completely unsupervised fashion. They experiment with segmenting books into chapters, which offers natural incidental supervision.

Another approach performs the segmentation task in a completely supervised manner, similar to supervised labeled span extraction tasks. At first, the models were LSTM-based (Arnold et al. 2019; Koshorek et al. 2018), and later on, Transformer-based (Lukasik et al. 2020; Somasundaran et al. 2020). Unlike finetuning, this approach requires large amounts of segmented data.

All of these works were designed and evaluated with structured written text, such as book chapters, Wikipedia pages, or artificially stitched segments, where supervised data is abundant. In this work, we address the segmentation of texts of which we have little supervised data regarding segment boundaries. We, therefore, adopt elements from the unsupervised approaches combined with supervised components and design a model for a novel segmentation task of unstructured spoken narratives.

3. Spoken Narratives: Holocaust Testimonies as a case Study

3.1 Corpus

Our data consists of Holocaust survivor testimonies. We received 1,000 testimonies from the Shoah Foundation. All testimonies were conducted orally with an interviewer, recorded on video, and transcribed as text. The lengths of the testimonies range from 2,609 to 88,105 words, with mean and median lengths of 23,536 and 21,424 words, respectively.

The testimonies were transcribed as time-stamped text. In addition, each testimony recording was divided into segments, typically a segment for each minute. Each segment was indexed with labels, possibly multiple. The labels are all taken from the SF thesaurus.² The thesaurus is highly detailed, containing ~ 8,000 unique labels across the segments. It's worth noting that the division into segments was done purely by length and does not take the labels into consideration.

3.2 Motivation for Topical Segmentation

Typically, narrative research faces a trade-off between the number of narrative texts, which is important for computational methods, and the specificity of the narrative context, which is essential for qualitative narrative research (Sultana et al. 2022). Holocaust testimonies provide a unique case of a large corpus with a specific context.

The large and specific corpus provides motivation for schema alignment. Assuming that there are common thematic units across testimonies, it should be possible to extract parallel segments. For example, many testimonies address a common event or experience (e.g., "deportation" or "physical hardship") and we might want to compare various aspects of the different reports. To do this we first need to align the testimonies according to these topics.

As opposed to some work that focuses on events in a narrative (Gius and Vauth 2022), we choose to focus on topics. This is for multiple reasons. First, our data is not completely event-based. There are parts that are not events (e.g., "family life", "reflection") and the focus is on the personal experience and not on historical facts. Second, event-complex extraction is a highly complex computational task and has very scarce annotated data (Ning et al. 2018).

2. See: <https://sfi.usc.edu/content/keyword-thesaurus>.

As opposed to traditional topic modeling, which is completely unsupervised, we decided to make use of the large annotated testimony corpus. We create a supervised dataset for topic classification and use it in a similar manner as a topic model.

3.3 Challenges

In spoken texts, there are no obvious boundaries. This is unlike common data used for segmentation, such as stitched-up texts (Misra et al. 2011), book chapters (Pethe et al. 2020), or Wikipedia sections (Arnold et al. 2019). Manual inspection of human-annotated segments in the testimonies shows that in many cases the segment boundary highly depends on the expected topic (see Figure 1). In some cases, there is clearly a topic change, but it is not clear where the transition took place. Even in cases where the boundary is in accordance to surface cues, the topics still play a role as they decide what should be read together.

We argue that proper segmentation cannot truly be separated from the topical structure. When there are no surface-form cues, the segments must depend on a higher-level structure. We claim that there is an essential difference between the prediction of existing segments, in which case the textual cues are given and play a heavy role in the creation of the segmentation, and the generation of a new segmentation for unstructured text, in which case the topical structure is more dominant.

Our experiments (see below) show that given annotated segments with topics, knowledge of the segment boundaries is highly beneficial for predicting the sequence of topics. Put differently, in the topical segmentation task, the segmentation boundaries are predictive of the topic sequence. This finding motivates the exploration of the task of topical segmentation even if the goal of the user is to extract a sequence of topics (and their localization in the text is of less utility).

The SF annotations use a very large set of labels. This hurts the uniformity of the annotation and raises computational issues. Therefore, it is necessary to reduce the size of the label set as we describe in the following section.

3.4 Building the Datasets

Classification. As some of the labels in the SF annotations are very rare, and given the noise in the data, using the full SF label set directly as classification labels is dispreferred. Instead, we reduced the number of labels through an iterative process of manual expert annotation and clustering. The SF thesaurus uses a hierarchical system of labels, ranging from high-level topics (e.g., “politics”, “religion and philosophy”), through mid-level (e.g., “camp experiences”, “ghetto experiences”), to low-level labels (e.g., “refugee camp injuries”, “forced march barter”). For the purpose of compiling the list of topics, we focused on mid-level labels. Then, with the help of domain experts from the field of Holocaust studies, we created a list of 29 topics that were deemed sufficiently informative, yet still generalizable across the testimonies. We added the label *NO-TOPIC*, which was used for segments that address technical details of the testimony-giving event (e.g., changing the tape), and do not include Holocaust-related content. For the full topic list see Table 1.

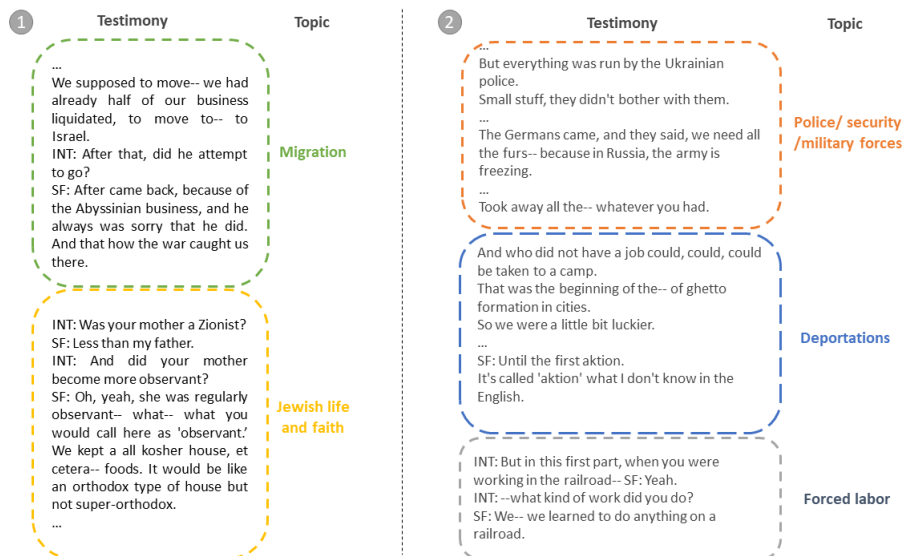


Figure 1: Examples of topic-related segment changes. Both examples are from SF testimony no. 43109.

No	Topic	Description
1	Adaptation and survival	Any act of finding ways to adapt to the war and persecution and to survive in Ghetto, camps, etc.
2	After the war	Not liberation, but post-war life
3	Aid	Either giving or receiving aid
4	Antisemitism and persecutions	This mostly refers to pre-war episodes, before the ghetto or camps
5	Before the war	This mostly refers to the opening parts relating the pre-war life in the hometown, family, friends, school, etc.
6	Betrayals	Any betrayal by friends, neighbors, locals, etc.
7	Brutality	Any acts of brutality, physical or mental, during the war – intended and performed by someone. To be distinguished from hardship which can describe a certain condition of hardship
8	Camp	Any events that take place in the concentration or death camps
9	Deportations	Deportation from the city / village to the ghetto, and from the ghetto to the camps. This includes any forced transport to an undesired destination
10	Enemy collaboration	Either Jews or locals collaborating with the Nazi regime or their representatives

11	Escape	Any escape from hometown, from the ghetto, from prison or camps
12	Extermination / execution / death march	Any event of violent intended killing
13	Extreme	Killing of a child, suicide, surviving a massacre
14	Family and friendships	Stories involving family members, friends, loved ones
15	Forced labor	Any events taking place in labor camps or as part of forced labor
16	Ghetto	Any event taking place in the ghetto
17	Hardship	Any description of physical or mental hardship
18	Hiding	Hiding places, woods, homes while running away or stories of being hidden by others (farms, monasteries, etc.)
19	Jewish life and death	Any event relating to Jewish life and its practices – school, prayer, shabbat, synagogue, before, during and after the war
20	Liberation	Events relating to allies' liberation of camps
21	Migration	Either pre- or post-war migration to other countries
22	Non Jewish faith	Any mention of non-Jewish beliefs, practices, etc.
23	Police / security / military forces	Events relating to soldiers and police, either enemy or allies
24	Political activity	Protests, political parties, either for or against Nazis
25	Prison	Captivity in prison – to be distinguished from camps
26	Reflection / memory / trauma	
27	Refugees	Mostly the post-war episodes in refugee / displaced persons camps
28	Resistance and partisans	Any act of resistance, organized or individual
29	Stills	Presentation of pictures

Table 1: List of topics with their description.

We filtered out testimonies that were not annotated in the same fashion as the others, for example, testimonies that did not have one-minute segments or ones that skipped segments altogether. We used these testimonies for development and testing. We also filtered out all segments that had more than one label after the label conversion. We ended up with a text classification dataset of 20,722 segments with 29 possible labels.

We added to the input texts an extra token to indicate the location within the testimony. We divided each testimony into 10 bins with equal segment counts and added the bin index to the input text.

Segmentation. The testimonies in the SF were divided based on pure length. Segment boundaries can appear in the middle of a topic or even a sentence. Therefore, the SF segments cannot be used for the evaluation of topical segmentation. Instead, for evaluation and test sets, we manually segmented and annotated 20 full testimonies. In this set of testimonies, the segmentation was performed based on a given list of topics. We used testimonies from SF that were not annotated in the same manner as the others, and therefore not seen for the classifier. The annotation was carried out by two trained annotators, highly proficient in English. We note that this process was done independently from the SF indexing and therefore the number of testimonies is relatively small.

An initial pilot study to segment testimonies without any prior requirements and no topic list yielded an approximate segment length (the results of these attempts were not included in the training or test data). The approximate length was not used as a strict constraint, but rather as a weak guideline, so as to align our expectations with the annotators.

The approximate desired average segment length was given to the annotators as well as the final topic list. The first annotator annotated all 20 testimonies, which were used for development and testing. The second annotator annotated 7 documents, used for measuring the inter-annotator agreement. The full annotation guidelines can be found in the Supplementary Material (§[subsection A.2](#)).

Altogether, for our test data, we obtained 20 testimonies composed of 1,179 segments with topics. The segment length ranges from 13 to 8,772 words, with a mean length of ~ 485 . We randomly selected 5 testimonies for parameter estimation, and the remaining 15 were used as a test set.

4. Topical Segmentation

We have a document X consisting of n sentences $x_1 \dots x_n$, which we consider atomic units. Our task is to find $k - 1$ boundary points, defining k segments, and k topics, where every consecutive pair of topics is different.

4.1 General Considerations

Designing a model for topical segmentation involves multiple, possibly independent, considerations which we present here.³

Our general approach to segmentation requires a scoring method that can be applied to each possible segment. Given these scores and the desired number of segments, we can then select the segmentation with the highest score.

3. For more technical details, Wagner et al. [2022](#).

We compose a segment score based on both local and non-local properties. For local scores, we propose to use Point-wise Mutual Information (PMI). Given a language model (LM), we hypothesize that the mutual information between two adjacent sentences can predict how likely the two sentences are to be in the same segment. These scores need additional supervision beyond the LM pretraining. Given these scores, the extraction of a segmentation for a given text is equivalent to maximizing the LM likelihood of text, under the assumption that each sentence depends on one previous sentence and that each segment depends on no previous sentences. A formal proof can be found in the supplementary material ([subsection A.1](#)).

This is opposed to recent work in this direction that uses the Next Sentence Prediction (NSP) scores (Pethe et al. 2020). We argue that the pretrained NSP scores do not capture the probability of two given consecutive sentences being in the same segment, since even if the second sentence is in another segment, it still is the next sentence.

Based on previous work, we also consider the non-local properties of segment length (Pethe et al. 2020), and topical coherence (Misra et al. 2011). Given a domain-specific multi-label classifier, we use the classification log probabilities as the coherence scores.

Given the desired number of segments, we have a structured prediction task that requires dynamic programming in order to be executed in polynomial time, where the degree of the polynomial is decided by the order of dependency. Inference of the optimal topic assignment according to a given classifier also requires a dynamic algorithm to avoid identical adjacent topics.

4.2 Topical Segmentation Models

We propose various models and baselines for the task of topical segmentation. Some models perform segmentation and topic assignment separately (“pipeline”) and some jointly.

Topic-Modeling Based. Misra et al. (2011) performed topical segmentation based on topic modeling, where the selected segmentation is that with the highest likelihood, based on the Latent Dirichlet Allocation model (LDA, Blei et al. 2003). The topic model gives a likelihood score to each segment and the segmentation that maximizes the product of likelihoods is selected. Inference is equivalent to finding the shortest path in a graph with n^2 nodes.

NSP-based Segmentation. The approach in the first ChapterCaptor model is to perform linear segmentation based on Next Sentence Prediction (NSP) scores. Using a model that was pretrained for NSP, they further finetune the model with segmented data, where a positive label is given to two subsequent spans in one segment, and a negative label is given to two spans that are in different segments.

The second ChapterCaptor model leverages the assumption that segments tend to have similar lengths. Given data, they compute the expected average length, L , and add regularization towards average-length segments. We denote this model with NSP+L.

LMPMI-based Segmentation. Adapting the NSP scores for segmentation seems sub-optimal in domains for which we do not have enough segmented data. We propose to replace the NSP scores with language-modeling (LM) and Point-wise Mutual Information (PMI) scores. Specifically, for each possible boundary index i , we define:

$$LMPMI_i = \log \frac{P_{LM}(x_i, x_{i+1})}{P_{LM}(x_i) \cdot P_{LM}(x_{i+1})} \quad (1)$$

where the probabilities are the LM probabilities for the sentences together or alone.

These scores can be computed by any pretrained language model, and the log-scores replace the NSP scores in both previous methods. We denote these models with PMI and PMI + L.

Pipeline Topic Assignment. Given a segmentation for the document and a topic classifier, we infer a list of topics. We need to find the optimal topic sequence under the constraint of no identical adjacent elements. This can be formalized as an HMM inference task, which can easily be found using dynamic programming.

Joint Segmentation and Topic Assignment. In another method, we take into account the segment classification scores in addition to a length penalty. We jointly infer a segmentation and topic assignment using the following dynamic formula:

$$\begin{aligned} cost(n, k, t) = \min_{\substack{1 \leq i \leq n-1 \\ t' \in T}} (cost(i, k-1, t') + \alpha \cdot \frac{|n-i-L|}{L} + \beta \cdot \log P(t'|X_i \dots X_n)) \\ + (1 - \alpha - \beta) \cdot PMI_n \quad (2) \end{aligned}$$

where $cost(n, k, t)$ represents the cost of a boundary at index n with $k-1$ previous boundaries and topic t as the last topic. α, β are hyperparameters controlling the components. We denote this model with PMI + T.

Baseline Models. As a point of comparison, we also implemented simple baseline models for segmentation and topic selection. These models can be used in a pipeline.

For segmentation, we divide a text into equally lengthed segments, given a predetermined number of segments. This method was used by Antoniak et al. (2019) and, with slight modifications, by Jeff Wu et al. (2021), as it is extremely simple and efficient.

For topic assignment, we sequentially sample topics from a uniform distribution over the set of given topics. We avoid repeating topics by giving probability 0 to the previous topic.

We denote these baselines with UNIFORM.

Implementation Specifics. The classifier was selected by fine-tuning various Transformer-based models with a classification head. Base models were pretrained by HuggingFace.⁴ We experimented with Distilbert, Distilroberta, Electra, RoBERTa, XLNet, and DeBERTa in various sizes. For our experiments we chose to use Distilroberta, which showed an accuracy score of ~ 0.55 , which was close to that of the larger models, doing this with way faster training and inference. We trained with a random 80-20 data split on 2 GPUs for ~ 10 minutes with the *Adam* optimizer for 5 epochs with *batch-size*=16, *label-smoothing*=0.01 and other settings set as default. We selected this classifier for our final segmentation experiments.

From the 20 manually segmented testimonies, we randomly took 5 testimonies a development set for hyperparameter tuning. Based on the results on this set, we chose $\alpha = 0.8$ for the PMI + L model and $\alpha = \beta = 0.2$ for the PMI + T model.

The LDA topic model was pretrained on the same training data as the classifier's (subsection 3.4), before running the segmentation algorithm. We trained the LDA model with 15 topics using the Gensim package,⁵ which we also used for the likelihood estimation of text spans given an LDA model.

We used HuggingFace's pretrained transformer models for the NSP scores and LM probabilities. We used FNET (Lee-Thorp et al. 2021) for NSP and GPT2 (Radford et al. 2019) for LM probabilities. We tuned the context size parameter on the development set, resulting in $C = 3$.

4.3 Evaluation Methods

Here we discuss appropriate metrics for the segmentation and topic assignments.

Segmentation. Measuring the quality of text segmentation is tricky. We want to give partial scores to segmentations that are close to the manually annotated ones, so simple Exact-Match evaluation is overly strict. This is heightened in cases like ours, where there is often no clear boundary for the topic changes. For example, in one place the witness says *"he helped us later when we planned the escape"*. This sentence comes between getting help (the *Aid* topic) and escaping (the *Escape* topic). We would like to give at least partial scores for boundaries either before or after this sentence.

Various attempts have been made to alleviate this problem and propose more relaxed measures. Since the notion of "closeness" strongly depends on underlying assumptions, it seems hard to pinpoint one specific measure that will perfectly fit our expectations. Following this rationale, we report a few different measures.

The first measure we report is the average F1 score, which counts overlaps in the exact boundary predictions. Another measure we used is average WindowDiff (WD; Pevzner and Hearst 2002), which compares the number of reference boundaries that fall in an interval with the number of boundaries that were produced by the algorithm. We also measured the average Segmentation Similarity (S-SIM; Fournier and Inkpen 2012) and Boundary Similarity (B-SIM; Fournier 2013) scores. These scores are based on the number of edits required between a proposed segmentation and the reference, where

4. See: <https://pypi.org/project/transformers/>.

5. See: <https://radimrehurek.com/gensim/>.

Boundary Similarity assigns different weights to different types of edits. In F1, B-SIM, and B-SIM a higher score is better and in WindowDiff a lower score is better. We used the segeval python package⁶ with the default settings to compute all of these measures. Notably, the window size was set to be the average segment length (in the reference segmentation for the particular testimony) divided by 2.

Topic Assignment. To measure the similarity between a predicted topic assignment and the reference assignment we used two different measures. One measure was python’s difflib SequenceMatcher (SM) scores, which are based on the gestalt pattern matching metric (Ratcliff and Metzener 1988), that considers common substrings. In this metric, a higher score means stronger similarity.

Another measure we used is the Damerau–Levenshtein edit distance (Edit, Damerau 1964), which measures distance by the number of actions needed to get from one sequence to another. We normalized the edits by the number of topics in the reference data. For the Edit distance, lower is better.

5. Results and Discussion

We evaluate our models for both the segmentation and the resulting topic sequence.

We do not report scores for the LDA-based model since it did not produce a reasonable number of segments, and its runtime was prohibitively long (in previous work, it was run on much shorter text). We also implemented the models with different sizes of GPT2. Observing that the size had no significant effect, we report the results with the base model (“gpt2”) only.

5.1 Annotator Agreement

Evaluating on the 7 documents that were annotated by both annotators, we achieve *Boundary score* = 0.324, *Sequence Matching* = 0.4 and *Edit distance* = 0.73.

In complex structured tasks, the global agreement score is expected to be low. Agreement in these cases is therefore often computed in terms of sub-structures (e.g., attachment score or PARSEVAL F-score in parsing instead of exact-match). Since no local scores are common in segmentation tasks, we report only the global scores despite their relative strictness. Compared to the boundary score of uniform-length segmentation (which is much better than random), we can see that the annotator agreement was larger by an order of magnitude. Eyeballing the differences between the annotators also revealed that their annotations are similar.

We note that the annotators did not always mark the same number of segments (and topics), and this can highly influence the scores. We also note that the annotators worked completely independently and did not adjudicate.

5.2 Model Performance

6. See: <https://pypi.org/project/segeval/>.

Model	F1	WD	S-SIM	B-SIM
UNIFORM	0.052	0.568	0.958	0.026
NSP + L	0.04	0.584	0.958	0.02
PMI	0.172	0.537	0.963	0.094
PMI + L	0.173	0.535	0.964	0.095
PMI + T	0.165	0.54	0.962	0.09

Table 2: Segmentation scores. We evaluate PMI-score models with and without length penalties (PMI and PMI + L, respectively). We also evaluate a joint model for segmentation with topics (PMI + T), a uniform length segmentor (UNIFORM) and a Next Sentence Prediction segmentor with length penalties (NSP + L). For F1, S-SIM and B-SIM, higher is better and for WD lower is better. The number of segments is decided using the expected segment length.

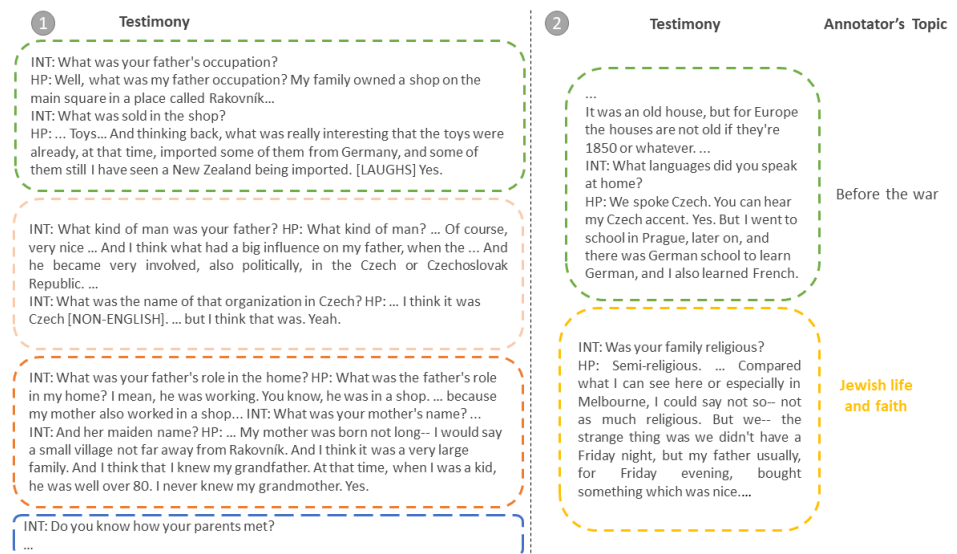


Figure 2: Examples of outputs from the PMI segmentation model. In 1 the predicted boundaries were not marked by the annotators. In 2 the model and the annotators agreed. Both examples are from SF testimony 43109.

Segmentation. Table 2 presents the results for the segmentation task. We see that PMI-based models are significantly better than the uniform length segmentation and the NSP-based model. Among the PMI-based models, there is no clear advantage for a specific setting, as the local PMI model is slightly better than the models with global scores.

We note that due to the nature of the metrics, specifically how they normalize the values to be between 0 and 1, the different measures vary in the significance of the gaps.

In Figure 2 we present two examples of outputs of the PMI model. In the first case, the human annotator did not put boundaries where the model did, but the model's predictions seem plausible. In the second example, the model predicted a boundary in the same place as the annotator.

Topic Assignment. Table 3 presents our results for the topic assignments produced by our models and the baselines. For comparison, it also presents the scores for topic creation based on the classifier when the real annotated segments are given.

Model	SM	Edit
UNIFORM	0.138	1.13
UNIFORM + CL	0.378	0.872
NSP + CL	0.369	0.875
PMI + CL	0.36	0.892
PMI + T	0.375	0.872
GOLD + CL	0.478	0.5

Table 3: Performance of the various models for topic lists. In Sequence Matching (SM) higher is better and for the Edit Distance, lower is better. In all cases, the number of topics was set as the length divided by the expected segment length rounded. The models we evaluate are uniform segmentation, NSP segmentation with length penalties, and PMI segmentation, all with dynamic topic assignment based on the classifier (UNIFORM + CL, NSP + CL and PMI + CL, respectively), and the joint segmentation and classification model (PMI + T). The baseline model is uniform topic generation (UNIFORM), which samples topics independently of the given text, and avoids repeating the previous topic.

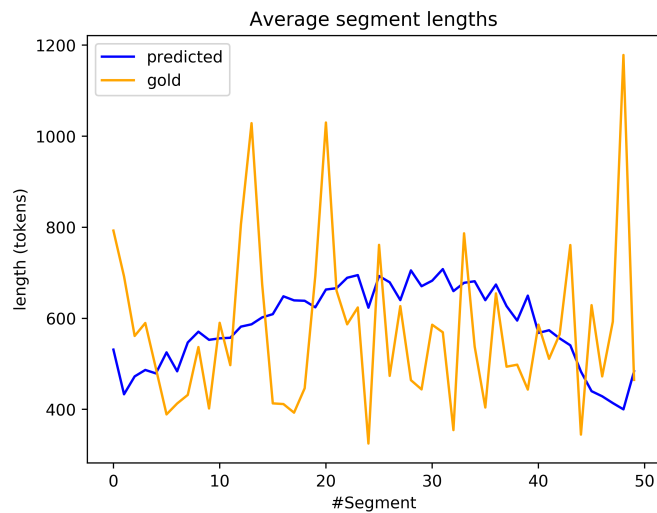


Figure 3: Segment length (in tokens) as a function of the segment index in the testimony. The predicted segments are decided by the PMI model for 50 segments. The gold segmentation was normalized to 50 segments and then averaged.

Here we see that the pipeline methods with uniform or NSP segmentation provide slightly better topics than the joint inference model or the simple PMI model. All models based on the classifier perform significantly better than the baselines.

5.3 Discussion

Our experiments show that topic assignment given the real segmentation (GOLD + CL) yields better topics than all other models. This suggests that a high-quality segmentation does contribute to the topic assignment, which motivates work on segmentation, even if the desired product is (only) the sequence of topics, without their localization. The GOLD + CL model in fact achieves higher topic similarity than the inter-annotator agreement. This might be explained by the fact that the GOLD + CL model was given the exact number of segments, while this was not specified for the annotators.

Regarding the segmentation models, our results show that the PMI methods present better performance for the segmentation task, compared to previous methods. However, we find that the automatic segmentation results do not contribute to the topic assignment. Also, within our different PMI models, we see that additional length and topic scores do not yield substantial improvements, neither for the segmentation nor for the topics. This is somewhat surprising and might mean that the sensitivity of our classifier to exact boundaries is low, or that the produced segments did not yet cross a usefulness threshold for topic classification.

Another seemingly surprising result is that larger sizes and domain fine-tuning of the GPT2 model do not improve performance, sometimes actually hurting it.

Delving more into the outputted segments created by the PMI models (Figure 3), it seems that these models do produce meaningful segments with good boundaries, but they do not always match the manual boundaries, as the exact segmentation also depends on the given set of topics. We hypothesize that there is a gap between the “surface level topic changes”, that are reflected in clear textual cues (e.g., a new question by the interviewer) and “high-level topic changes”, that highly depend on our prior domain-specific topical interests.

If the hypothesis is correct we would expect to see more PMI boundaries in parts of the testimony that are more structured, compared to the reference boundaries that would be more prevalent where there are more Holocaust-related themes.

In Figure 3 we plot the average lengths (in tokens) of the segments as a function of the location within the testimony. Since the number of segments in the reference data varies, we normalize the lengths as if there were 50 segments in each testimony. We can see that the PMI segments tend to be longer in the middle of the testimony and shorter at the beginning and end. The reference segmentation has less of a pattern.⁷ This result supports our hypothesis. The SF testimonies have a relatively uniform structure at the beginning and the end, so it should be easier to detect surface-level changes there.

This is an important argument regarding contemporary segmentation models. It is common to test the model with highly structured cases, like book sections (Pethe et al. 2020), Wikipedia sections (Arnold et al. 2019), and randomly stitched stories (Misra et al. 2011). In these cases, it is important to assert that the high performance is not due to surface-level cues, in which case the model only predicts traces of previously generated sections and does not actually segment a long document.

We note that the data for the classification and segmentation are restricted to a specific domain, specifically Holocaust testimonies. This limits the generalization of our models to other domains. The general ideas are domain-independent, and some models can be readily used, the application of the models that use the classifier will require adaptation to a new domain.

We also note that the use of single-label topic classification has its limitations. It seems that in some cases the topics are not mutually exclusive (e.g., a segment can involve both “Family” and “Ghetto”). This makes the topical segmentation task less conclusive. In

7. We averaged over 500 predicted testimonies and only 20 reference ones, so it is expected that the reference lengths will be noisier.

future work we intend to model the temporal and spatial paths of testimonies, allowing segmentation and alignment in a more robust manner.

6. Conclusion

We presented models for combined segmentation and topic extraction for narratives. We found that: (1) segmentation boundaries can be indicative of the sequence of topics (as demonstrated by using the gold standard segmentation; however, (2) topic lists inferred dynamically given a classifier are not very sensitive to the actual segmentation, allowing the extraction of high-quality topic lists even with uniform segmentation. In addition, we find that (3) local PMI scores are sufficient to infer a segmentation with better quality than previous models; (4) additional features such as segment lengths and topics seem to have limited influence on the quality of the segmentation;

Our work addresses the segmentation and topic labeling of text in a naturalistic domain, involving unstructured, transcribed text. Our model can segment noisy texts where textual cues are sparse.

In addition to the technical contribution of this work, it also makes important first steps in analyzing spoken testimonies in a systematic, yet ethical manner. With the imminent passing of the last remaining Holocaust survivors, it is increasingly important to design methods of exploration and analysis of these testimonies, so as to enable us to use the wealth of materials collected in the archives for studying and remembering their stories.

7. Ethical Statement

We abided by the instructions provided by each of the archives. We note that the witnesses identified themselves by name, and so the testimonies are not anonymous. Still, we do not present in the analysis here any details that may disclose the identity of the witnesses. We intend to release our codebase and scripts, but those will not include any of the data received from the archives; the data and trained models used in this work will not be given to a third party without the consent of the relevant archives. We note that we did not edit the testimony texts in any way. The compilation of the Holocaust related topics was done by Holocaust experts based on the SF thesaurus hierarchy. We did not apply automatic generation at any point.

8. Data Availability

Data will be given upon permission from the Shoah Foundation.

9. Software Availability

Software can be found here: <https://github.com/eitanwagner/holocaust-segmentation>

10. Acknowledgements

The authors acknowledge the USC Shoah Foundation – The Institute for Visual History and Education for its support of this research. We thank Prof. Gal Elidan, Prof. Todd Presner, Dr. Gabriel Stanovsky, Gal Patel and Itamar Trainin for their valuable insights and Nicole Gruber, Yelena Lizuk, Noam Maeir and Noam Shlomai for research assistance. This research was supported by grants from the Israeli Ministry of Science and Technology and the Council for Higher Education and the Alfred Landecker Foundation.

11. Author Contributions

Eitan Wagner: Conceptualization, Investigation, Computational analysis, Visualization, Experimentation, Implementation, Writing – original draft, Writing – review and editing

Renana Keydar: Conceptualization, Data Curation, Supervision, Writing – original draft, Writing – review and editing

Amit Pinchevski: Conceptualization, Writing – original draft, Writing – review and editing

Omri Abend: Conceptualization, Supervision, Writing – original draft, Writing – review and editing

References

- Alhussain, Arwa I. and Aqil M. Azmi (2021). “Automatic Story Generation: A Survey of Approaches”. In: *ACM Computing Surveys* 54 (5), 1–38. [10.1145/3453156](https://doi.org/10.1145/3453156).
- Antoniak, Maria, David Mimno, and Karen Levy (2019). “Narrative Paths and Negotiation of Power in Birth Stories”. In: *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW), 1–27. [0.1145/3359190](https://doi.org/10.1145/3359190).
- Arnold, Sebastian, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser (2019). “SECTOR: A Neural Model for Coherent Topic Segmentation and Classification”. In: *Transactions of the Association for Computational Linguistics* 7, 169–184. [10.1162/tacl_a_00261](https://doi.org/10.1162/tacl_a_00261).
- Artstein, Ron, Alesia Gainer, Kallirroi Georgila, Anton Leuski, Ari Shapiro, and David Traum (2016). “New Dimensions in Testimony Demonstration”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Association for Computational Linguistics, 32–36. [10.18653/v1/N16-3007](https://doi.org/10.18653/v1/N16-3007).
- Bamman, David, Brendan O’Connor, and Noah A. Smith (2013). “Learning Latent Personas of Film Characters”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 352–361. <https://aclanthology.org/P13-1035> (visited on 11/29/2023).
- Bamman, David, Ted Underwood, and Noah A. Smith (2014). “A Bayesian Mixed Effects Model of Literary Character”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 370–379. [10.3115/v1/P14-1035](https://doi.org/10.3115/v1/P14-1035).

- Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, 993–1022. <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (visited on 12/05/2023).
- Castricato, Louis, Stella Biderman, David Thue, and Rogelio Cardona-Rivera (2021). "Towards a Model-Theoretic View of Narratives". In: *Proceedings of the Third Workshop on Narrative Understanding*. Association for Computational Linguistics, 95–104. [10.18653/v1/2021.nuse-1.10](https://aclanthology.org/2021.nuse-1.10).
- Chambers, Nathanael (2013). "Event Schema Induction with a Probabilistic Entity-Driven Model". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1797–1807. <https://aclanthology.org/D13-1185> (visited on 11/29/2023).
- Chambers, Nathanael and Dan Jurafsky (2009). "Unsupervised Learning of Narrative Schemas and their Participants". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics, 602–610. <https://aclanthology.org/P09-1068> (visited on 11/29/2023).
- Damerau, Fred J. (1964). "A Technique for Computer Detection and Correction of Spelling Errors". In: *Communications of the ACM* 7 (3), 171–176. [10.1145/363958.363994](https://doi.org/10.1145/363958.363994).
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint*. [10.48550/ARXIV.1810.04805](https://arxiv.org/abs/1810.04805).
- Du, Lan, John Pate, and Mark Johnson (2015). "Topic Segmentation with an Ordering-Based Topic Model". In: 29. [10.1609/aaai.v29i1.9502](https://arxiv.org/abs/1609.04805).
- Eisenstein, Jacob and Regina Barzilay (2008). "Bayesian Unsupervised Topic Segmentation". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 334–343. <https://aclanthology.org/D08-1035> (visited on 11/29/2023).
- Fogu, Claudio, Wulf Kansteiner, and Todd Presner (2016). *Probing the Ethics of Holocaust Culture*. Harvard University Press.
- Fournier, Chris (2013). "Evaluating Text Segmentation using Boundary Edit Distance". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1702–1712. <https://aclanthology.org/P13-1167> (visited on 11/29/2023).
- Fournier, Chris and Diana Inkpen (2012). "Segmentation Similarity and Agreement". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 152–161. <https://aclanthology.org/N12-1016> (visited on 11/29/2023).
- Gius, Evelyn and Michael Vauth (2022). "Towards an Event Based Plot Model. A Computational Narratology Approach". In: *Journal of Computational Literary Studies* 1 (1). [10.48694/jcls.110](https://doi.org/10.48694/jcls.110).
- He, Xin, Jian Wang, Quan Zhang, and Xiaoming Ju (2020). "Improvement of Text Segmentation TextTiling Algorithm". In: 1453. [10.1088/1742-6596/1453/1/012008](https://doi.org/10.1088/1742-6596/1453/1/012008).
- Hearst, Marti A. (1997). "Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages". In: *Computational linguistics* 23 (1), 33–64. [10.5555/972684.972687](https://doi.org/10.5555/972684.972687).

- Jeong, Minwoo and Ivan Titov (2010). “Multi-document Topic Segmentation”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, 1119–1128. [10.1145/1871437.1871579](https://doi.org/10.1145/1871437.1871579).
- Jockers, Matthew L. and David Mimno (2013). “Significant Themes in 19th-century Literature”. In: *Poetics* 41 (6), 750–769. [10.1016/j.poetic.2013.08.005](https://doi.org/10.1016/j.poetic.2013.08.005).
- Kazantseva, Anna and Stan Szpakowicz (2012). “Topical Segmentation: a Study of Human Performance and a New Measure of Quality.” In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 211–220. <https://aclanthology.org/N12-1022> (visited on 11/29/2023).
- (2014). “Hierarchical Topical Segmentation with Affinity Propagation”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, 37–47. <https://aclanthology.org/C14-1005> (visited on 11/29/2023).
- Keydar, Renana (2019). “Mass Atrocity, Mass Testimony, and the Quantitative Turn in International Law”. In: *Law & Society Review* 53 (2), 554–587.
- Keydar, Renana, Yael Litmanovitz, Badi Hasisi, and Yoav Kan-Tor (2022). “Modeling Repressive Policing: Computational Analysis of Protocols from the Israeli State Commission of Inquiry into the October 2000 Events”. In: *Law & Social Inquiry* 4, 1075–1105. [10.1017/lsi.2021.63](https://doi.org/10.1017/lsi.2021.63).
- Koshorek, Omri, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant (2018). “Text Segmentation as a Supervised Learning Task”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 469–473. [10.18653/v1/N18-2075](https://doi.org/10.18653/v1/N18-2075).
- Lee-Thorp, James, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon (2021). “Fnet: Mixing Tokens with Fourier Transforms”. In: *arXiv preprint*. [10.48550/arXiv.2105.03824](https://arxiv.org/abs/10.48550/arXiv.2105.03824).
- Li, Manling, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss (2020). “Connecting the Dots: Event Graph Schema Induction with Path Language Modeling”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 684–695. [10.18653/v1/2020.emnlp-main.50](https://doi.org/10.18653/v1/2020.emnlp-main.50).
- Li, Wenbo, Tetsu Matsukawa, Hiroto Saigo, and Einoshin Suzuki (2020). “Context-Aware Latent Dirichlet Allocation for Topic Segmentation”. In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan. Springer International Publishing, 475–486. [10.1007/978-3-030-47426-3_37](https://doi.org/10.1007/978-3-030-47426-3_37).
- Lukasik, Michal, Boris Dadachev, Kishore Papineni, and Gonçalo Simões (2020). “Text Segmentation by Cross Segment Attention”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 4707–4716. [10.18653/v1/2020.emnlp-main.380](https://doi.org/10.18653/v1/2020.emnlp-main.380).
- Mikhalkova, Elena, Timofei Protasov, Polina Sokolova, Anastasiia Bashmakova, and Anastasiia Drozdova (2020). “Modelling Narrative Elements in a Short Story: A Study on Annotation Schemes and Guidelines”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 126–132. <https://aclanthology.org/2020.lrec-1.16> (visited on 11/29/2023).

- Min, Semi and Juyong Park (2019). “Modeling Narrative Structure and Dynamics with Networks, Sentiment Analysis, and Topic Modeling”. In: *PLOS ONE* 14 (12). [10.1371/journal.pone.0226025](https://doi.org/10.1371/journal.pone.0226025).
- Misra, Hemant, François Yvon, Olivier Cappé, and Joemon Jose (2011). “Text Segmentation: A Topic Modeling Perspective”. In: *Information Processing and Management* 47 (4), 528–544. [10.1016/j.ipm.2010.11.008](https://doi.org/10.1016/j.ipm.2010.11.008).
- Mota, Pedro, Maxine Eskenazi, and Luísa Coheur (2019). “BeamSeg: A Joint Model for Multi-Document Segmentation and Topic Identification”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 582–592. [10.18653/v1/K19-1054](https://doi.org/10.18653/v1/K19-1054).
- Ning, Qiang, Hao Wu, and Dan Roth (2018). “A Multi-Axis Annotation Scheme for Event Temporal Relations”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1318–1328. [10.18653/v1/P18-1122](https://doi.org/10.18653/v1/P18-1122).
- Pethe, Charuta, Allen Kim, and Steve Skiena (2020). “Chapter Captor: Text Segmentation in Novels”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 8373–8383. [10.18653/v1/2020.emnlp-main.672](https://doi.org/10.18653/v1/2020.emnlp-main.672).
- Pevzner, Lev and Marti A. Hearst (2002). “A Critique and Improvement of an Evaluation Metric for Text Segmentation”. In: *Computational Linguistics* 28 (1), 19–36. [10.1162/089120102317341756](https://doi.org/10.1162/089120102317341756).
- Piper, Andrew, Richard Jean So, and David Bamman (2021). “Narrative Theory for Computational Narrative Understanding”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 298–311. [10.18653/v1/2021.emnlp-main.26](https://doi.org/10.18653/v1/2021.emnlp-main.26).
- Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). “Language Models are Unsupervised Multitask Learners”. In: *OpenAI blog* 1 (8), 9.
- Rashkin, Hannah, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao (2020). “PlotMachines: Outline-Conditioned Generation with Dynamic Plot State Tracking”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 4274–4295. [10.18653/v1/2020.emnlp-main.349](https://doi.org/10.18653/v1/2020.emnlp-main.349).
- Ratcliff, John W. and David E. Metzener (1988). “Pattern Matching: The Gestalt Approach”. In: *Dr Dobbs Journal* 13 (7), 46.
- Reiter, Nils (2015). “Towards Annotating Narrative Segments”. In: *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Association for Computational Linguistics, 34–38. [10.18653/v1/W15-3705](https://doi.org/10.18653/v1/W15-3705).
- Reiter, Nils, Judith Sieker, Svenja Guhr, Evelyn Gius, and Sina Zarriß (2022). “Exploring Text Recombination for Automatic Narrative Level Detection”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, 3346–3353. <https://aclanthology.org/2022.lrec-1.357> (visited on 11/29/2023).
- Riedl, Martin and Chris Biemann (2012). “TopicTiling: A Text Segmentation Algorithm based on LDA”. In: *Proceedings of ACL 2012 Student Research Workshop*. Association for

- Computational Linguistics, 37–42. <https://aclanthology.org/W12-3307> (visited on 11/29/2023).
- Schöch, Christof (2017). “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama”. In: *Digital Humanities Quarterly* 11 (2). <https://www.digitalthumanities.org/dhq/vol/11/2/000291/000291.html> (visited on 12/05/2023).
- Shaham, Uri, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy (2022). “SCROLLS: Standardized CompaRison Over Long Language Sequences”. In: *arXiv preprint*. [10.48550/arXiv.2201.03533](https://arxiv.org/abs/10.48550/arXiv.2201.03533).
- Somasundaran, Swapna et al. (2020). “Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 34, 7797–7804.
- Sultana, Sharifa, Renwen Zhang, Hajin Lim, and Maria Antoniak (2022). “Narrative Datasets through the Lenses of NLP and HCI”. In: ed. by Su Lin Blodgett, Hal Daumé III, Michael Madaio, Ani Nenkova, Brendan O’Connor, Hanna Wallach, and Qian Yang, 47–54. [10.18653/v1/2022.hcinlp-1.7](https://arxiv.org/abs/10.18653/v1/2022.hcinlp-1.7).
- Uglanova, Inna and Evelyn Gius (2020). “The Order of Things. A Study on Topic Modelling of Literary Texts”. In: *CHR 18-20, 2020*. <https://ceur-ws.org/Vol-2723/long7.pdf> (visited on 12/05/2023).
- Wagner, Eitan, Renana Keydar, Amit Pinchevski, and Omri Abend (2022). “Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 6809–6821. [10.18653/v1/2022.emnlp-main.457](https://arxiv.org/abs/10.18653/v1/2022.emnlp-main.457).
- Wang, Haoyu, Hongming Zhang, Muhao Chen, and Dan Roth (2021). “Learning Constraints and Descriptive Segmentation for Subevent Detection”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5216–5226. [10.18653/v1/2021.emnlp-main.423](https://arxiv.org/abs/10.18653/v1/2021.emnlp-main.423).
- Wilmot, David and Frank Keller (2020). “Modelling Suspense in Short Stories as Uncertainty Reduction over Neural Representation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1763–1788. [10.18653/v1/2020.acl-main.161](https://arxiv.org/abs/10.18653/v1/2020.acl-main.161).
- (2021). “A Temporal Variational Model for Story Generation”. In: *arXiv preprint*. [10.48550/ARXIV.2109.06807](https://arxiv.org/abs/10.48550/ARXIV.2109.06807).
- Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano (2021). “Recursively Summarizing Books with Human Feedback”. In: *arXiv preprint*. [10.48550/arXiv.2109.10862](https://arxiv.org/abs/10.48550/arXiv.2109.10862).
- Wu, Qiong, Adam Hare, Sirui Wang, Yuwei Tu, Zhenming Liu, Christopher G Brinton, and Yanhua Li (2020). “BATS: A Spectral Biclustering Approach to Single Document Topic Modeling and Segmentation”. In: *arXiv preprint*. [10.48550/arXiv.2008.02218](https://arxiv.org/abs/10.48550/arXiv.2008.02218).
- Zehe, Albin, Leonard Konle, Lea Katharina Dümpelmann, Evelyn Gius, Andreas Hotho, Fotis Jannidis, Lucas Kaufmann, Markus Krug, Frank Puppe, Nils Reiter, Annekea Schreiber, and Nathalie Wiedmer (2021). “Detecting Scenes in Fiction: A new Segmentation Task”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, 3167–3177. [10.18653/v1/2021.eacl-main.276](https://arxiv.org/abs/10.18653/v1/2021.eacl-main.276).

Zhai, Fangzhou, Vera Demberg, Pavel Shkadzko, Wei Shi, and Asad Sayeed (2019). "A Hybrid Model for Globally Coherent Story Generation". In: *Proceedings of the Second Workshop on Storytelling*. Association for Computational Linguistics, 34–45. [10.18653/v1/W19-3404](https://doi.org/10.18653/v1/W19-3404).

A. Appendix: Supplementary Material

A.1 Equivalence of PMI and Likelihood

We have a document $X = x_1, x_2 \dots, x_n$ which we want to divide into k segments.

We assume that the LM probability for each sentence depends only on the previous sentence and that in the case of a boundary at index i , sentence i is independent of all previous sentences. Under these assumptions, the segmentation that places boundaries at the places with minimal PMI is the same segmentation that maximized the LM likelihood.

Proof: Assume we have a boundary set $B = (i_1, i_2, \dots, i_k)$.

For any $i \in B$ we have:

$$PMI(x_i, x_{i-1}) = \frac{P(x_i|x_{i-1})}{P(x_i)} = 1$$

Therefore we get:

$$\begin{aligned} {}_B P(X) &= {}_B P(X) \cdot \prod_{i=1}^n \frac{1}{P(x_i)} = {}_B \prod_{i \notin B} \frac{P(x_i|x_{i-1})}{P(x_i)} \prod_{i \in B} \frac{P(x_i)}{P(x_i)} \\ &= {}_B \sum_{i \notin B} \log PMI(x_i, x_{i-1}) = {}_B \sum_{i=1}^n \log PMI(x_i, x_{i-1}) \quad (3) \end{aligned}$$

A.2 Annotation Guidelines

Annotation Guidelines for Topical Segmentation

In this task, we divide Holocaust testimonies into topically coherent segments. The topics for the testimonies were predetermined. We have 29 content topics and a NULL topic. The full list is attached. Each segment has one topic (multi-class, not multi-label), and a change of topic is equivalent to a change of segments.

The segmentation annotation will be as follows:

- The testimonies are already divided into sentences. A segment change can only be between sentences.
- Our goal is to annotate segmentations. For this, we will assign a topic for each sentence. Since the main focus is the segment, the topic should be given based on a segment and not a single sentence.
- The changing of a topic, if it does not include further information, should not be marked as a separate topic, rather it should be combined with the surrounding topics. If there is a change of topics there then the Overlap should be marked as True over these sentences.

- Regarding the number of requested segments, we want an approximate average segment length of 30 sentences. This is a global attribute, as the actual Segment lengths can (and should) vary, depending on the topics. Any single segment should be decided mainly by content and not by constraints regarding the segment lengths.
- After deciding the segment scope, all sentences can be marked at once. No need to mark them one by one.
- No sentence should be left without a topic (NULL is also a topic). If the topic is unclear then one should be chosen. It should not be left empty.
- A “thumb rule” in cases of multiple options is to choose a topic that is more Holocaust-specific. For example, a hiding story about a family member should be assigned to “Hiding” and not to “Family and Friendships”.