Topical Segmentation of Spoken Narratives: A Test Case on Holocaust Survivor Testimonies

Eitan Wagner[†] **Renana Keydar**[‡] **Amit Pinchevski**[◇] **Omri Abend**[†]
[†] Department of Computer Science [‡] Faculty of Law and Digital Humanities
[◇] Department of Communication and Journalism
Hebrew University of Jerusalem
{first_name}.{last_name}@mail.huji.ac.il

Abstract

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. 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 explore a range of algorithmic approaches to the task, building on previous work on segmentation that uses generative Bayesian modeling and state-of-the-art neural machinery. Compared to manually annotated references, we find that the developed approaches show considerable improvements over previous work.[1]

1 Introduction

Proper representation of narratives in long texts remains an open problem in NLP (Piper et al., 2021; Castricato et al., 2021; Mikhalkova et al., 2020). 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). A common modern approach for modeling narratives is as a sequence of neural states (Wilmot and Keller, 2020, 2021; Rashkin et al., 2020). However, a drawback of this approach is the lack of interpretability, which is crucial in some contexts.

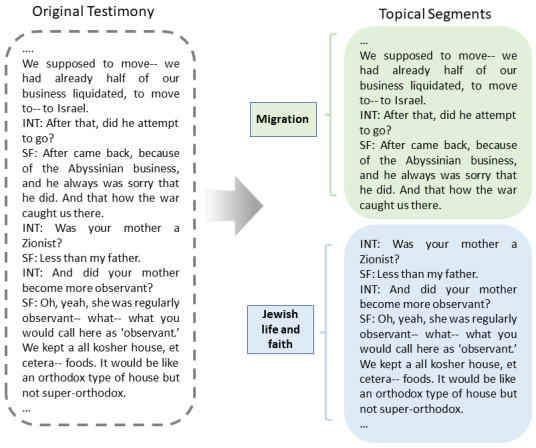

Figure 1: Topical segmentation in Holocaust testimonies.

A different approach represents and visualizes a narrative as a sequence of interpretable topics (Antoniak et al., 2019). 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. 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 (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 don't have large datasets of divided text. Moreover, there may not be any obvious boundaries that can be derived

[1] Code is provided at https://github.com/eitanwagner/holocaust-segmentation.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6809–6821
December 7-11, 2022

based on local properties. This makes the task more challenging and hampers the possibility of taking a completely supervised approach.

We propose an unsupervised alternative for segmentation, based on two assumptions: (1) segment boundaries correspond to places with low mutual information between sentences over the boundary; (2) neural language models can serve as reliable sentence probability estimators. Based on these assumptions, we propose a simple approach to segmentation and offer extensions involving dynamic programming. The proposed models give a substantial margin over the existing methods in terms of segmentation performance. In order 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).[2] Inspired by Misra et al. (2011), we also incorporate the topical coherence based on the topic classifier into the segmentation model.

Our contributions are the following: (1) we present the task of topical segmentation for running, unedited text; (2) we propose novel algorithmic methods for tackling the task without any manual segmentation supervision, building on recent advances in language modeling; (3) comparing to previous work, we find substantial improvements over existing methods; (4) we compile a test set for evaluation in the case of Holocaust testimonies; (5) we develop domain-specific topical classifiers to extract lists of topics for long texts.

Typically, narrative research faces a tradeoff 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. Our work also communicates with Holocaust research, seeking methods to better access testimonies as the survivor generation is slowly passing away (Artstein et al., 2016). We expect our methods to promote schema-based analysis and browsing of testimonies, enabling better access and understanding.

2 Previous work

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 (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 (Li et al., 2020b).

Following recent advances in neural models, these models have been used for the task of supervised text segmentation. Pethe et al. (2020) introduced 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 towards 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 (Koshorek et al., 2018; Arnold et al., 2019), and later on, Transformer based (Somasundaran et al., 2020; Lukasik et al., 2020). Unlike finetuning, this approach requires a large amount of segmented data.

All of these works were designed and evaluated with structured written text, such as book chapters,

[2] https://sfi.usc.edu/

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.

Narrative analysis. Much work has been done in the direction of probabilistic schema inference, focusing on either event schemas (Chambers and Jurafsky, 2009; Chambers, 2013; Li et al., 2020a) or persona schemas (Bamman et al., 2013, 2014).

Recently, neural models were utilized for story modeling. Wilmot and Keller (2020) presented a neural GPT2-based model for suspense in short stories. This work follows an information-based framework, modeling the reader's suspense by different types of predictability. Due to their strong performance in text generation, neural models are commonly used for story generation, with numerous structural variations (Zhai et al., 2019; Rashkin et al., 2020; Alhussain and Azmi, 2021).

Narrative analysis can help in conveying the essence of stories, without all the details. This can aid the meta-analysis of stories. 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, using simplistic, uniform segmentation with topic modeling to visualize the frequent topic paths.

3 Methods

We have a document X consisting of n sentences $x_1 \cdots x_n$, which we consider as 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.

3.1 Design Principles of Used Methods

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

Local Potential-Boundary Scores. A simple approach to text segmentation involves giving independent local scores to each possible boundary. Given these scores and the desired number of segments, we can then select the best boundaries.

Recent work in this direction uses the Next Sentence Prediction (NSP) scores (Pethe et al., 2020). Given two sentences x_1, x_2, their NSP score is defined as the predicted probability that the second sentence actually came after the first and not from somewhere else. The prediction is usually carried out using a pretrained model with a self-supervised training protocol and is typically further finetuned for a specific task.

We argue that the pretrained NSP scores do not capture the probability of two given sequential sentences being in the same segment, since even if the second sentence is in a new segment, it still is the next sentence. Therefore, we expect this approach to perform poorly in settings for which there are not enough segmented texts for finetuning.

Instead, we propose to use Point-wise Mutual Information (PMI) for the local boundary scores. 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 assumptions that each sentence depends on one previous sentence, and that each segment depends on no previous sentences (for proof see Appendix A).

Non-local Scores. Full segmentation of text involves the selection of multiple boundaries, and these selections might not be independent. Even a single segment directly involves two boundaries. Therefore, we might want to use scores that take into account properties that involve more than one boundary. Given scores for all possible segments, we can optimize for the maximal total score over all possible segmentations.

A simple property that was used in previous work is the segment length (Pethe et al., 2020), with a higher score given to segments whose length is closer to the expected length. These scores can be helpful if we assume that segments' length tends to be close to uniform. These scores can also be used in a conditional manner, in case we have estimates for the segment lengths of different topics or in different locations of the whole text. Segment length scores require the consideration of at least two corresponding boundaries for each score.

Another property that was used in previous work is topic scores (Misra et al., 2011). Given some Topic Model (TM), we can use the generation log-likelihood of a segment as its score. Alternatively,

with supervised data for multi-label classification, we can use the classification log probabilities. With these scores, we can optimize for the maximal sum.

Since we assume adjacent segments to have different topics, these scores must consider at least 3 boundaries creating two adjacent segments.

Pipeline, Joint Inference, Independent inference. The task of topical segmentation involves the extraction of both a segmentation and a corresponding topic assignment for a given document. We consider 3 options for the inferential setup: (1) sequential inference, where we first infer a segmentation and then derive a topic assignment given the segmentation ("pipeline"); (2) joint inference, where we jointly optimize for the segmentation and the topic assignment; and (3) independent inference, where we infer the segments regardless of the topics, and the topics regardless of the segments.

In addition, topical segmentation requires a number of segments k. This can be decided in a pipeline (i.e., first decide k) or jointly (i.e., infer k together with the boundaries).

Independent and pipeline inference are generally less complex algorithmically, as they allow decomposition of the problem.

Local Decoding vs. Dynamic Programming. Given a desired number of segments, and considering only local scores, we can easily select the optimal segmentation in one linear pass. If we also consider global scores then 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.

Given a segmentation, the inference of the optimal topic assignment might require dynamic programming. Since we require adjacent segments to have different topics, greedy local topic inference might not give the optimal topic assignment.

3.2 Models

We propose various models and baselines for the task of topical segmentation. Each model is defined as a combination of the possibilities listed above.

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

This method jointly infers the number of segments, the segmentation, and topical distributions for each segment. Fixing the number of segments ahead of time requires complex inference as it adds a restriction on the segmentation.[3]

NSP. 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 spans can be single sentences or some fixed length window. The NSP score is the probability given to the positive label. We denote the NSP score for placing a boundary point after the n-th sentence as NSP_n.

After computing the NSP score for all $1 \leq i \leq n$, the segmentation is derived by selecting the k places where the NSP scores are lowest and placing boundaries there. We denote this model with NSP.

NSP with length penalties. 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.

Specifically, they use the dynamic formula:

$$cost(n,k) = \min_{1 \leq i \leq n-1} \left(cost(i,k-1) + (1-\alpha)\frac{|n-i-L|}{L}\right) + \alpha \cdot NSP_n \quad (1)$$

where $cost(n,k)$ represents the cost of putting a boundary at index n when we already have $k-1$ previous boundaries. α is a hyperparameter controlling the balance between the two factors.

We denote this model with *NSP + L*.

LMPMI. 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

[3] Misra et al. (2011) mention that they used a penalty factor for the number of segments, but it remains unclear how it was actually used in the framework, as it introduces dependencies between segment boundaries.

(LM) and Point-wise Mutual Information (PMI) scores. Specifically, for each possible boundary index i, we define:

$$LMPMI_i = \frac{P_{LM}(x_i, x_{i+1})}{P_{LM}(x_i) \cdot P_{LM}(x_{i+1})} \quad (2)$$

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.

3.3 Topic Assignment

Pipeline. Given a segmentation for the document and a topic classifier, we can infer a list of topics. We need to find the optimal topic sequence under the constraint of no identical adjacent elements.

Finding the optimal topic assignment can be formalized as an HMM inference task, which can easily be found using dynamic programming. Assuming uniform prior probabilities for the topics, the initial state probability is uniform and the transition probabilities are uniform over all states other than the current one. The trained classifier gives us the probabilities of a topic given a segment, $P(t|X)$. With our assumption of uniform topic probabilities, these probabilities are proportional to the emission probabilities.

Joint Inference. As an extension to the previous methods, we propose a formula that takes into account the segment classification scores in addition to the lengths. This is based on the assumption (similar to Misra et al. (2011)), that topically coherent segments will have classification probabilities that are concentrated around the best topic.

Using these scores, we can jointly infer a segmentation and topic assignment. We use the following dynamic formula:

$$\begin{aligned} &cost(n, k, t) = \\ &\min_{\substack{1 \le i \le n-1 \\ t' \in T}} \big(cost(i, k-1, t') + \alpha \cdot \frac{|n - i - L|}{L} \\ &+ \beta \cdot \log P(t'|X_i \cdots X_n)\big) + (1 - \alpha - \beta) \cdot PMI_n \end{aligned} \quad (3)$$

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.

3.4 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.

Uniform Segmentation. The simplest way to segment a text is to divide it into equally lengthed segments, given a predetermined number of segments. This method was used by Antoniak et al. (2019) and, with slight modifications, by Wu et al. (2021), as it is extremely simple and efficient. We set k as the specific document length divided by the average number of tokens per segment in the development set. This baseline is denoted with UNIFORM.

Uniform Topic Selection. Given the length of the topic list to extract for the text, we can sequentially sample topics from a uniform distribution over the set of topics. In this case, we can easily avoid repeating topics by giving probability 0 to the previous topic. This too is denoted with UNIFORM.

4 Experimental Setup

4.1 Data

Our data consists of Holocaust survivor testimonies. We received 1000 testimonies from SF. All testimonies were conducted orally with an interviewer, recorded on video, and transcribed as text. The lengths of the testimonies range from 2609 to 88105 words, with a mean length of 23536 words.

Data for the Classifier. The testimonies, originally recorded, 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.[4] The thesaurus is highly detailed, containing $\sim$ 8000 unique labels across the segments.

As some of the labels are very rare, and given the noise in the data, using the full label set directly 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 labels (e.g., "camp experiences", "ghetto experiences"),

[4] https://sfi.usc.edu/content/keyword-thesaurus

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. (the full list can be found in Appendix C).

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 20722 segments with 29 possible labels.

Since the segments were determined based on time intervals and not content, we cannot use this data as supervision for boundaries, as was done in recent work on segmentation.

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 number to the input text.

Test Data for Segmentation. To compile evaluation and test sets for the topical segmentation problem, we manually segmented and annotated 20 testimonies. We used testimonies from SF that were not annotated in the same manner as the others, and therefore not used for the classifier. The annotation was carried out by two trained annotators, highly proficient in English.

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 just 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 Appendix B.

Altogether, for our test data, we obtained 20 testimonies composed of 1179 segments with topics. The segment-length ranges from 13 to 8772 words, with a mean length of $\sim$ 485. We randomly selected 5 testimonies for parameter estimation, and the remaining 15 were used as a test set.

4.2 Classifier Specifics

The classifier was selected by fine-tuning various Transformer-based models with a classification head. Base models were pretrained by HuggingFace.[5] 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 $\sim$ 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 $\sim$ 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.

We also experimented with a number of linear classifiers. The highest test accuracy we achieved was $\sim$ 0.46, which is considerably lower than the one achieved with the neural classifiers. We, therefore, did not use any of the linear models in the final segmentation experiments.

4.3 Model Specifics

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 (§4.1), before running the segmentation algorithm. We trained the LDA model with 15 topics using the Gensim package,[6] 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 experimented with different context sizes C (i.e., how many sentences on each side we use

[5] https://pypi.org/project/transformers/

[6] https://radimrehurek.com/gensim/

for comparison). We tuned the size parameter on the development set, resulting in $C = 3$.

With this setting, the dynamic model with topics takes approximately 50 minutes per testimony, the dynamic model without topics takes approximately 5 minutes per testimony, and the simple gpt2 model takes approximately 2 minutes per testimony, all running with 1 GPU.

4.4 Evaluation Methods

Evaluating the classifier component is straightforward since we have labeled data and we can use a held-out test set. We note that the classifier was trained on data that was not divided by topic. We report accuracy scores. 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 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 package[7] 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. One measure we used was python's difflib SequenceMatcher (SM) scores, which are based on the *gestalt pattern matching* metric (Ratcliff and Metzener, 1988). This metric sums the longest common substrings in a recursive manner, and divides by the total length, attempting to reflect human impression for similarity. In this metric, a higher score means stronger similarity.

Another measure we used is the Damerau–Levenshtein edit distance (Edit, Damerau 1964). This measure defines the distance between two sequences as the minimal number of insertions, deletions, substitutions, or transpositions in order to get from one sequence to the other. Since the number of edits depends on the number of elements in the sequence, we normalized the distance by the number of topics in the reference document.[8] For the Edit distance, lower is better.

5 Results

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.

We emphasize that our models are only weakly supervised, as the topic classifier was trained with arbitrary boundaries and the topics were implicitly derived from the data.

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

[7] https://pypi.org/project/segeval/

[8] We can still get a distance larger than 1 if the predicted number of topics is larger than the real number. This normalization is commonly known in the literature as *word error rate*.

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 1: 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.

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.

Segmentation. Table 1 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.

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. S-SIM normalizes by all possible boundaries, so the score will always be high since in most places there is no boundary, even for low quality segmentations. In fact, the cosmetically high values of S-SIM were one of the incentives for the definition of B-SIM (Fournier, 2013). WD uses a sliding window. Therefore, it essentially normalizes by the number of possible boundaries, but, unlike B-SIM, WD usually counts errors multiple times, resulting in lower scores.

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 2: 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 topics 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.

Topic lists. Table 2 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.

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.

6 Discussion

Our results show that topic assignment given the real segmentation GOLD + CL gives better topics than all other models. This suggests that a good segmentation does contribute to the topic assignment, which motivates tackling the segmentation and topic assignment jointly, in principle. The GOLD + CL model actually 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 full models, our results show that the PMI methods show better performance for the segmentation task, compared to previous methods. This supports our hypothesis that segment boundaries correspond to low mutual information between the segments. This connects to common

unsupervised methods and trends, showing that general-purpose self-supervised models can perform strongly on various tasks.

However, we find that the automatic segmentation results do not contribute to the topic assignment (topic assignment scores are comparable with uniform segmentation). It seems then that although the PMI methods show improvement over previous work for the segmentation task, their results are still not sufficient to contribute to the topic extraction.

Within the 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 surprising result is that larger sizes and domain fine-tuning of the GPT2 model do not improve the performance, sometimes actually hurting it. Inspecting the produced segments, it seems that these models do produce meaningful segments with good boundaries, but they don't always match the manual boundaries, as the exact segmentation depends also on the given set of topics. That said, the fact that the models still perform better than baseline models shows that it is possible to produce reasonable segmentations even without specifying a set of topics.

7 Conclusion

We presented models for combined segmentation and topic extraction for narratives. We found that: (1) local PMI scores are sufficient to infer a segmentation with better quality than previous models; (2) additional features such as segment lengths and topics seem to have limited influence on the quality of the segmentation; (3) 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.

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 browsing and analyzing these testimonies, so as to enable us to use the wealth of materials collected in the archives for studying and remembering the stories.

Limitations

Our data for the classification and segmentation are restricted to a specific domain. This limits the generalization of our models to other domains. This is true both regarding the application of the models, as models that use the classifier will require adaptation to a new domain, and regarding the results of the experiments.

Another limitation regards the task of segmentation, as it is not always defined in the same manner and depends on the specific requirements in place. This is true for both supervised and unsupervised methods, since even within a specific domain the optimal segmentation may vary.

Ethical Considerations

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.

Acknowledgments

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.

References

Arwa I. Alhussain and Aqil M. Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Comput. Surv.*, 54(5).

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27.

Sebastian Arnold, Rudolf Schneider, Philippe Cudré-Mauroux, Felix A. Gers, and Alexander Löser. 2019. SECTOR: A neural model for coherent topic segmentation and classification. *Transactions of the Association for Computational Linguistics*, 7:169–184.

Ron Artstein, 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*, pages 32–36, San Diego, California. Association for Computational Linguistics.

David Bamman, 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)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman, 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)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Louis Castricato, 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*, pages 95–104, Virtual. Association for Computational Linguistics.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA. Association for Computational Linguistics.

Nathanael Chambers 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*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176.

Lan Du, John K Pate, and Mark Johnson. 2015. Topic segmentation with an ordering-based topic model. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 334–343, Honolulu, Hawaii. Association for Computational Linguistics.

Chris Fournier. 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)*, pages 1702–1712, Sofia, Bulgaria. Association for Computational Linguistics.

Chris Fournier 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*, pages 152–161, Montréal, Canada. Association for Computational Linguistics.

Xin He, Jian Wang, Quan Zhang, and Xiaoming Ju. 2020. Improvement of text segmentation texttiling algorithm. In *Journal of Physics: Conference Series*, volume 1453, page 012008. IOP Publishing.

Marti A Hearst. 1997. Text tiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Minwoo Jeong and Ivan Titov. 2010. Multi-document topic segmentation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1119–1128.

Omri Koshorek, 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)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020a. 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)*, pages 684–695, Online. Association for Computational Linguistics.

Wenbo Li, Tetsu Matsukawa, Hiroto Saigo, and Einoshin Suzuki. 2020b. Context-aware latent dirichlet allocation for topic segmentation. *Advances in Knowledge Discovery and Data Mining*, 12084:475.

Michal Lukasik, 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)*, pages 4707–4716, Online. Association for Computational Linguistics.

Elena Mikhalkova, 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*, pages 126–132, Marseille, France. European Language Resources Association.

Semi Min and Juyong Park. 2019. Modeling narrative structure and dynamics with networks, sentiment analysis, and topic modeling. *PloS one*, 14(12):e0226025.

Hemant Misra, François Yvon, Olivier Cappé, and Joemon Jose. 2011. Text segmentation: A topic modeling perspective. *Information Processing & Management*, 47(4):528–544.

Pedro Mota, 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)*, pages 582–592, Hong Kong, China. Association for Computational Linguistics.

Charuta Pethe, 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)*, pages 8373–8383, Online. Association for Computational Linguistics.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, 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)*, pages 4274–4295, Online. Association for Computational Linguistics.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Martin Riedl and Chris Biemann. 2012. TopicTiling: A text segmentation algorithm based on LDA. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42, Jeju Island, Korea. Association for Computational Linguistics.

Uri Shaham, 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.

Swapna Somasundaran et al. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7797–7804.

Sharifa Sultana, Renwen Zhang, Hajin Lim, and Maria Antoniak. 2022. Narrative datasets through the lenses of nlp and hci.

Haoyu Wang, 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*, pages 5216–5226, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Wilmot 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*, pages 1763–1788, Online. Association for Computational Linguistics.

David Wilmot and Frank Keller. 2021. A temporal variational model for story generation. *CoRR*, abs/2109.06807.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nissan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Qiong Wu, 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. *arXiv preprint arXiv:2008.02218*.

Fangzhou Zhai, 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*, pages 34–45, Florence, Italy. Association for Computational Linguistics.

A Equivalence of PMI and Likelihood

We have a document $X = x_1, x_2 \cdots, x_n$ 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 $B = (i_1, i_2, ...i_k)$.

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:

$$\begin{aligned}
\arg\max_B P(X) &= \arg\max_B P(X) \cdot \prod_{i=1}^{n} \frac{1}{P(x_i)} \\
&= \arg\max_B \prod_{i \notin B} \frac{P(x_i|x_{i-1})}{P(x_i)} \prod_{i \in B} \frac{P(x_i)}{P(x_i)} \\
&= \arg\max_B \sum_{i \notin B} logPMI(x_i, x_{i-1}) \\
&= \arg\max_B \sum_{i=1}^{n} logPMI(x_i, x_{i-1}) \quad (4)
\end{aligned}$$

B 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 tape, 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 ("None" 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".

C Topic list

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 of 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 faith	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 or resistance, organized or individual
29 Stills	Presentation of pictures