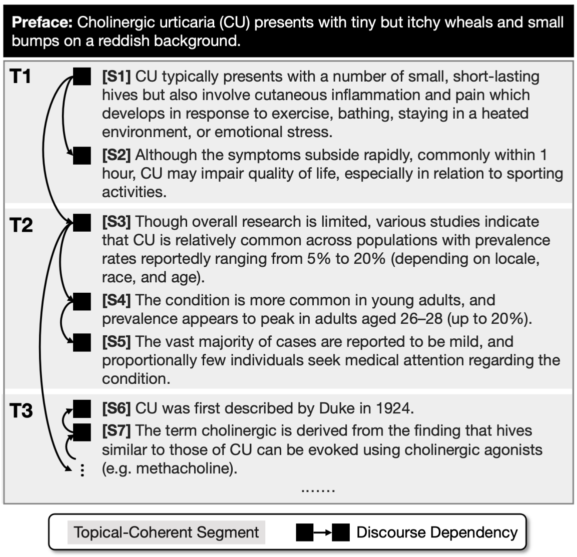
Improving Topic Segmentation by Injecting Discourse Dependencies

Linzi Xing , Patrick Huber , Giuseppe Carenini
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada, V6T 1Z4
{lzxing, huberpat, carenini}@cs.ubc.ca

Abstract

Recent neural supervised topic segmentation models achieve distinguished superior effectiveness over unsupervised methods, with the availability of large-scale training corpora sampled from *Wikipedia*. These models may, however, suffer from limited robustness and transferability caused by exploiting simple linguistic cues for prediction, but overlooking more important inter-sentential topical consistency. To address this issue, we present a discourse-aware neural topic segmentation model with the injection of above-sentence discourse dependency structures to encourage the model make topic boundary prediction based more on the topical consistency between sentences. Our empirical study on English evaluation datasets shows that injecting above-sentence discourse structures to a neural topic segmenter with our proposed strategy can substantially improve its performances on intra-domain and out-of-domain data, with little increase of model's complexity.



Preface: Cholinergic urticaria (CU) presents with tiny but itchy wheals and small bumps on a reddish background.

[S1] CU typically presents with a number of small, short-lasting hives but also involve cutaneous inflammation and pain which develops in response to exercise, bathing, staying in a heated environment, or emotional stress.

[S2] Although the symptoms subside rapidly, commonly within 1 hour, CU may impair quality of life, especially in relation to sporting activities.

[S3] Though overall research is limited, various studies indicate that CU is relatively common across populations with prevalence rates reportedly ranging from 5% to 20% (depending on locale, race, and age).

[S4] The condition is more common in young adults, and prevalence appears to peak in adults aged 26–28 (up to 20%).

[S5] The vast majority of cases are reported to be mild, and proportionally few individuals seek medical attention regarding the condition.

[S6] CU was first described by Duke in 1924.

[S7] The term cholinergic is derived from the finding that hives similar to those of CU can be evoked using cholinergic agonists (e.g. methacholine).

....

Figure 1: An example article about Cholinergic Urticaria (CU) sampled from the *en_disease* portion of Wiki-Section dataset (Arnold et al., 2019). Left: discourse dependency structure predicted by the Sent-First discourse parser (Zhou and Feng, 2022).

1 Introduction

Topic segmentation is a fundamental NLP task with the goal to separate textual documents into coherent segments (consisting of one or more sentences), following the document's underlying topical structure. The structural knowledge obtained from topic segmentation has been shown to play a vital role in key NLP downstream tasks, such as document summarization (Mitra et al., 1997; Riedl and Biemann, 2012; Xiao and Carenini, 2019), question answering (Oh et al., 2007; Diefenbach et al., 2018) and dialogue modeling (Xu et al., 2021; Zhang et al., 2020). The aim of topic segmentation makes it tightly connected to related research areas aiming to understand the latent structure of long and potentially complex text. Specifically, understanding the semantic and pragmatic underpinnings of a document can arguably support the task of separating continuous text into topical segments. To this end, discourse analysis and discourse parsing provide the means to understand and infer the semantic and pragmatic relationships underlying complete documents, well aligned with the local text coherence and highly correlated to the inter-sentential topical consistency, as shown in Louis and Nenkova (2012) and Muangkammuen et al. (2020). With a variety of linguistic theories proposed in the past, such as the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), the lexicalized discourse framework (Webber et al., 2003a) (underlying PDTB), and the Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Asher et al., 2003), we follow the RST framework in this work (1) as we focus on monologue text (as compared to dialogue frameworks, such as SDRT) and (2) since RST postulates complete discourse trees spanning whole documents, directly aligned with the topical structure of complete documents (Huber et al., 2021).

Proceedings of 3rd Workshop on Computational Approaches to Discourse, pages 7–18
October 16 – 17, 2022. ©2022 International Conference on Computational Linguistics

We further motivate the synergistic relationship between topic segmentation and discourse analysis/parsing in Figure 1, showing anecdotal evidence of the alignment between the document's topical structure and the respective RST-style discourse dependency graph. Starting from a sequence of sentences, the task of topic segmentation addresses the problem of splitting the given *Wikipedia* article into an ordered set of topical-coherent fragments (here: `T1`, `T2` and `T3`) by predicting topical boundaries. As shown in the example, the document discourse tree is indicative of the topical structure of the document, as discourse dependencies occur considerably more often within a topic segment than across topic segments.

Given significant influence on a variety of real-world tasks, topic segmentation is an active research area in the field of NLP. As such, modern, neural methods for monologue topic segmentation are proposed by formulating the task as a sentence-level sequence labeling problem, trained and evaluated on the large-scale *Wikipedia* dataset (Xing et al., 2020; Glavas and Somasundaran, 2020; Barrow et al., 2020; Lo et al., 2021). These Wikipedia articles are well-suited for the task of topic segmentation, providing natural section marks which can be reasonably used as ground-truth segment boundaries (Koshorek et al., 2018; Arnold et al., 2019), superseding previously proposed unsupervised methods (Hearst, 1997; Galley et al., 2003; Eisenstein and Barzilay, 2008; Song et al., 2016). Despite the significant improvements achieved by neural supervised topic segmentation models, it remains unclear if these topic segmenters effectively learn to cluster sentences into topical-coherent pieces based on the (document-level) topical consistency, or solely exploit superficial patterns (e.g., simple linguistic cues) in the training domain.

To address this challenge, in this paper, we propose a more discourse-aware neural topic segmentation model. We thereby inject above-sentence discourse structures into basic topic segmenter to encourage the model to base its topic boundary prediction more explicitly on the topical consistency between sentences. More specifically, we propose to exploit a discourse dependency parser pre-trained on out-of-domain data to induce inter-sentential discourse dependency trees. Subsequently, we convert the dependency tree into a directed discourse graph with sentences as nodes and discourse dependencies as edges. With the generated discourse graph, a Graph Attention Network (GAT) (Veličković et al., 2018) is used to encode sentences as discourse-contextualized representations by aggregating information from neighboring sentence nodes in the graph. Finally, the discourse-infused sentence representations are concatenated with standard encodings for segment boundary prediction.

In our empirical study conducted on English evaluation datasets, we show that: (i) Injecting discourse structures can substantially improve the performance of the basic neural topic segmentation model on three datasets. (ii) Our novel, discourse-enhanced topic segmenter is more robust compared to the basic neural model in settings that require domain transfer, showing superior performance on four challenging real-world test sets, to confirm the improved domain-independence. (iii) Even if our proposal has inferior accuracy against a state-of-the-art segmenter sharing the same basic architecture, it does achieve significantly better efficiency assessed by model's parameter size and speeds for learning and inference, which makes it potentially more favorable in real-world use.

2 Related Work

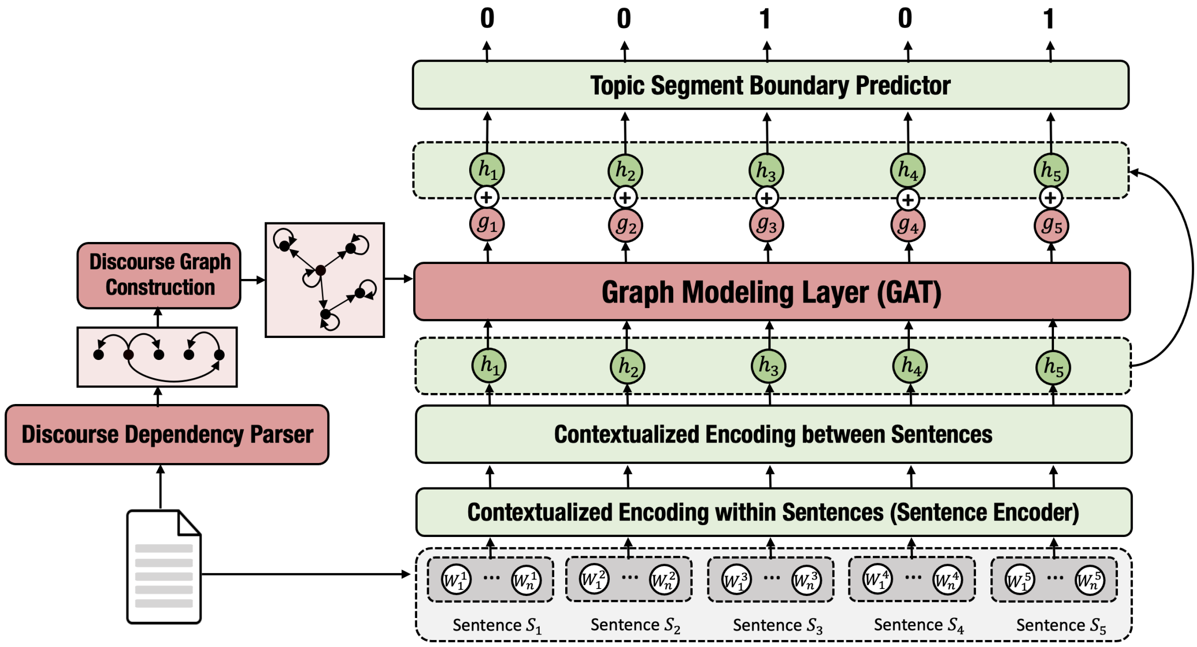
Topic Segmentation aims to reveal important aspects of the semantic structure of a document by splitting a sequence of sentences into topic-coherent textual units. Typically, computational topic segmentation models can be broadly separated into supervised and unsupervised approaches. Early topic segmentation methods usually fall into the category of unsupervised approaches, mainly due to the prevalent data sparsity issue at the time. Based on predicting the coherence between sentences through shallow (surface-level) features, unsupervised models reach a limited understanding of the contextualized structure of documents by merely relying on easy-to-extract but barely effective features for the similarity measurement between sentences (i.e., the degree of token overlap between two sentences) (Hearst, 1997; Eisenstein and Barzilay, 2008). Improving on the unsupervised topic segmentation paradigm, researchers started to address this issue by introducing pre-trained neural language models (LMs), trained on massive dataset (Xu et al., 2021; Solbiati et al., 2021; Xing and Carenini, 2021). Some works show that the signal captured in pre-trained LMs (e.g., BERT (Devlin et al., 2019)) are more indicative of topic relevance between sentences than early surface-level features. However, these proposed strategies of integrating BERT into the topic segmentation framework solely exploit BERT to induce dense encodings and further compute reciprocal sentence similarities. While this constitues a reasonable first step, the considerable gap between the training objective of LMs and topic segmentation task requires further efforts along this line of work (Sun et al., 2022).

More recently, the data sparsity issue has been alleviated by the proposal of large-scale corpora sampled from *Wikipedia* (e.g., Wiki-727k (Koshorek et al., 2018) and Wiki-Section (Arnold et al., 2019)), in which well-structured articles with their section marks are used as gold labels for segment boundaries. As a result, neural supervised topic segmenters started to gain attention by reaching greater effectiveness and efficiency compared to previously proposed unsupervised approaches. These supervised topic segmenters typically follow a common strategy which formulates the task as a sentence-level sequence labeling problem. More specifically, by assigning binary labels to each sentence, models infer the likelihood of a sentence to be a topic segment boundary (Koshorek et al., 2018; Arnold et al., 2019; Barrow et al., 2020; Lo et al., 2021). However, we believe that current models, besides reaching promising performance, potentially favour simple linguistic cues over effective measurements for semantic cohesion, restricting their application to narrow domains. Some recent works have attempted to address this limitation via explicitly integrating coherence modeling components into segmenters (Xing et al., 2020; Glavas and Somasundaran, 2020). However, compared to our objective in this work, these proposed coherence modeling strategies are either (i) only taking two adjacent sentences into account, limiting the additional module to extremely local contexts, or (ii) discriminating real documents from artificially “incoherent" texts, resulting in implicit and synthetic negative training samples and heavy parameter size caused by modeling multiple tasks simultaneously.

In contrast, we propose an effective method to integrate the document discourse (dependency) structure into neural topic segmentation frameworks, following the intuition that above-sentence discourse structure are indicative of text coherence and topical consistency, providing a more global and interpretable source of information for better topic transition prediction.

Discourse Analysis and Parsing analyze and generalize the underlying semantic and pragmatic structure of a coherence document (called a discourse). As an important upstream task in the field of NLP, discourse analysis proposes elaborate frameworks and theories to describe the textual organization of a document. To this end, a variety of popular discourse theories proposed in the past, such as (besides others) the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and the lexicalized discourse framework (Webber et al., 2003b) for monologues as well as the Segmented Discourse Representation Theory (SDRT) (Asher, 1993; Asher et al., 2003) for dialogues. Among these theories, the RST discourse theory postulates a single, complete discourse tree for monologue documents, while the lexicalized discourse framework only focuses on local discourse connectives within and between adjacent sentences. Focusing on the connection between discourse information and topic segmentation, we employ the RST discourse theory in this work, most aligned with the requirement to capture topical coherence.

Building on human annotated discourse treebanks, a mix of traditional and neural discourse parsers have been proposed over the last decades, with traditional approaches mainly exploiting surface-level features through Support-Vector Machines (SVMs) (Hernault et al., 2010; Ji and Eisenstein, 2014; Wang et al., 2017) or Conditional Random Fields (CRFs) (Joty et al., 2015; Feng and Hirst, 2014). On the other hand, neural models achieve similar or superior results on RST discourse parsing, with models using either custom architectures (Yu et al., 2018; Liu and Lapata, 2018) or pre-trained LMs (e.g. BERT (Zhou and Feng, 2022), RoBERTa (Guz et al., 2020), SpanBERT (Guz and Carenini, 2020)). In this work, we generate discourse dependency trees from a BERT-based neural dependency parser proposed in Zhou and Feng (2022), since: (i) The parser follows the intuition that information, and hence structures, in sentences are oftentimes “self-contained”. Therefore, it predicts the interactions between EDUs of the same sentence in a first stage and subsequently predicts the inter-sentential discourse structures, which aligns well with our objective of sentence-level topic segmenation. (ii) The parser by Zhou and Feng (2022) makes direct prediction of dependency discourse structures, alleviating the potential error caused by converting constituency structures



0
0
1
0
1
Topic Segment Boundary Predictor
h_1 h_2 h_3 h_4 h_5
g_1 g_2 g_3 g_4 g_5
Discourse Graph Construction
Graph Modeling Layer (GAT)
Contextualized Encoding between Sentences
Discourse Dependency Parser
Contextualized Encoding within Sentences (Sentence Encoder)
Sentence S_1
Sentence S_2
Sentence S_3
Sentence S_4
Sentence S_5

Figure 2: The overall architecture of our discourse-infused topic segmentation model.

into their respective dependency trees.

3 Methodology

As shown in Figure 2, our proposed discourse-aware neural topic segmentation model comprises two components: the *Hierarchical Topic Segmenter* and *Discourse Graph Modeling*, highlighted in green and red respectively. Discourse Graph Modeling further comprises of a *Discourse Graph Construction* and *Graph Modeling* component.

3.1 Basic Model: Hierarchical Topic Segmenter

The basic architecture of our proposal is adopted from the basic model in Xing et al. (2020), consisting of two hierarchical layers: First, a sentence encoder contextualizes individual sentences, followed by the second layer, conditioning sentences on the complete document. Following the settings in Xing et al. (2020), we adopt the attention BiLSTM architecture[1] for each layer and enhance the encodings with pre-trained BERT embeddings. Formally, given a document D as a sequence of n sentences, the sentence encoder (bottom component in Figure 2) yields the embedding for each individual sentence. Based on the obtained encodings, the document-level contextualization layer returns an ordered set of hidden states $\boldsymbol{H} = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_n\}$. Next, a simple multilayer perceptron (MLP) with a final softmax activation serves as a binary topic boundary predictor based on a threshold τ, tuned on the validation set. During training, we optimize the model in accordance to the cross-entropy loss, while at inference time, every sentence (except the last sentence[2]) with a probability $\geq \tau$ is considered as the end of a segment.

3.2 Discourse Graph Modeling

Our goal is to inject inter-sentential discourse dependency structures into the task of topic segmentation. We believe that the additional, structural information is thereby well aligned with the topical consistency between sentences, hence, suited to guide the prediction of topic transitions. To integrate the discourse information into the basic model described in section 3.1, we first generate an above-sentence discourse dependency tree $\boldsymbol{T}_D$ for the document. Specifically, we utilize the discourse dependency parsing model proposed in Zhou and Feng (2022), reaching state-of-the-art performance for discourse tree construction and relation type identification in multiple language settings. The "Sent-First" parser (Zhou and Feng, 2022) further fits the aim of our proposal due to its two-staged approach, first generating discourse trees within

[1] We also considered Transformer as the backbone of contextualized encoder, but eventually chose BiLSTM for its superior performance.

[2] We remove the last sentence from the sequence for prediction since it is per definition the end of the last segment.

Treebank	# of doc	# sent/doc	# edu/doc
RST-DT	385	22.5	56.6
GUM	150	49.3	114.2
SciDTB	1,355	5.3	14.1
COVID19-DTB	300	7.8	20.0

Table 1: Key dataset statistics of the discourse treebanks used for retraining the Sent-First discourse parser (Zhou and Feng, 2022).

sentences and subsequently combining sentence-level sub-trees. This hard constraint allows us to exclusively obtain above-sentence discourse structures, avoiding potentially leaky sub-trees (Joty et al., 2015). Regarding the discourse relations attached to every head-dependent pair (discourse dependency), we follow the observation in Xu et al. (2020), stating that the agreement between the type of rhetorical relation is usually lower and more ambiguous, to leave them for future work to avoid error propagation.

In contrast to the original proposal in Zhou and Feng (2022), training and testing their dependency discourse parser on one corpus (i.e., SciDTB (Yang and Li, 2018)), we believe that a mixture of several diverse and publicly available discourse treebanks with different document lengths and domains can increase the parser's robustness on new and unseen genres. Therefore, we retrain the parser on a mixture of RST-DT[3] (Carlson et al., 2002), GUM[4] (Zeldes, 2017), SciDTB[5] (Yang and Li, 2018) and COVID19-DTB[6] (Nishida and Matsumoto, 2022). More specifically, we combine those discourse treebanks and randomly split the aggregated corpus into 80% training, 10% validation, 10% test data. The parser retrained on our combined training portion achieves an Unlabeled Attachment Score (UAS) of 58.6 on the test portion. We show additional key dataset statistics for each treebank used in this paper in Table 1.

After training the discourse parser to infer a discourse dependency tree $\boldsymbol{T}_D$ for document D, we convert the tree structure into a discourse graph $\boldsymbol{G}_D$ (as a binary matrix). Formally, we initialize the graph $\boldsymbol{G}_D$ as a $n \times n$ identity matrix $\boldsymbol{G}_D = \boldsymbol{I}_{n,n}$, connecting every node to itself. Afterwards, we fill in the remaining cells by assigning $\boldsymbol{G}_D[i][j] = 1$ iff $\exists\ \boldsymbol{T}_D(i \rightarrow j)$, with i, j indexing the head and dependant sentences in the document, respectively. Using the binary matrix representation of $\boldsymbol{G}_D$, we apply the multi-layer Graph Attention Network (GAT) (Veličković et al., 2018) to update sentence encodings following the discourse graph. More specifically, with the discourse graph matrix $\boldsymbol{G}_D$ and the contextualized representations $\boldsymbol{H} = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_n\}$ described in section 3.1, within each graph attentional layer, we perform self-attention on the sentence nodes. Taking the lth layer as an example, we compute the attention coefficient $\alpha_{i,j}$ between sentence nodes i, j as:

$$\alpha_{ij}^l = softmax(e_{ij}^l) = \frac{exp(e_{ij}^l)}{\sum_{k \in \mathcal{N}_i} exp(e_{ik}^l)}, \quad (1)$$

$$e_{ij}^l = LeakyReLU(\boldsymbol{a}_l^T[\boldsymbol{W}_l \boldsymbol{g}_i^l || \boldsymbol{W}_l \boldsymbol{g}_j^l]) \quad (2)$$

where $\boldsymbol{W}_l$ and $\boldsymbol{a}_l$ are learnable parameters for layer l and ^{T} is the transposition operation. $\mathcal{N}_i$ denotes the direct neighborhood of node i in the graph ($G_D[i][\cdot] = 1$). As the node representation input of the first GAT layer ($l = 0$), $\boldsymbol{g}_i^0 = \boldsymbol{h}_i \in H$. Once attention coefficients are obtained, we compute the intermediate node representation $\boldsymbol{z}_i^l$ for sentence node i at layer l by aggregating information from neighboring nodes as:

$$\boldsymbol{z}_i^l = \sum_{j \in \mathcal{N}_i} \alpha_{ij}^l \boldsymbol{W}_l \boldsymbol{g}_j^l \quad (3)$$

Following the step in Huang et al. (2020), we combine the intermediate node representation $\boldsymbol{z}_i^l$ with the input of this layer $\boldsymbol{g}_i^l$ to get the updated node representation $\boldsymbol{g}_i^{l+1}$ as the input for the next layer:

$$\boldsymbol{g}_i^{l+1} = ELU(\boldsymbol{g}_i^l + \boldsymbol{z}_i^l) \quad (4)$$

where ELU denotes an exponential linear unit (Clevert et al., 2016). With the output $\boldsymbol{g}_i$ from the last layer of GAT, we concatenate it together with $\boldsymbol{h}_i$ and further feed $[\boldsymbol{h}_i; \boldsymbol{g}_i]$ into the predictor layer for segment boundary prediction.

4 Experiments

In order to quantitatively evaluate the effectiveness, generality and efficiency of our proposal, we conduct three sets of experiments to compare our topic segmentation approach against a variety of baselines and previous models. Namely, we assess the performance of our model in regards to the *Intra-Domain Segment Inference Performance*, *Domain Transfer Segment Inference Performance*, and conduct an additional *Efficiency Analysis*.

[3] catalog.ldc.upenn.edu/LDC2002T07

[4] corpling.uis.georgetown.edu/gum

[5] https://github.com/PKUTANGENT/SciDTB

[6] https://github.com/norikinishida/biomedical-discourse-treebanks

Dataset	# of doc	# sent/seg	# seg/doc
CHOI	920	7.4	10.0
RULES	4,461	7.4	16.0
SECTION	21,376	7.2	7.9

Table 2: Statistics of the datasets used in intra-domain experiments.

4.1 Datasets

4.1.1 Intra-Domain Datasets

For the set of intra-domain segment inference experiments, we train and test models within the same domain (here: on the same corpus). We thereby choose three diverse corpora (see Table 2 for more details) for the intra-domain evaluation:

Choi (Choi, 2000). This corpus consists of 920 articles artificially generated by randomly combining passages from the Brown corpus. The datapoints in this dataset are not human written, leading us to solely use this corpus for a preliminary performance assessment for topic segmentation models in a 80% (train)/10%(dev)/10%(test) data-split.

Rules (Bertrand et al., 2018). This corpus consists of 4,461 documents about regulation discussion published in the Federal Register[7] by U.S. federal agencies. Since each paragraph is about one particular regulation and all regulations covered by one document are under the same category, we deem it as a reasonably coherent data source for topic segmentation evaluation with the paragraph breaks as ground-truth segment boundaries. We split this dataset into training, validation and test sets with the default 80%, 10%, 10% data-split.

Wiki-Section (Section) (Arnold et al., 2019). This corpus originally contains Wikipedia articles in both English and German. The English portion of the dataset, which we use for our intra-domain experiment, consists of around 3.6k articles about diseases and 19.5k articles about cities around the world. After the step of filtering out problematic samples with incorrect sentence segmentation detected by mismatched counts between sentences and labels, the resulted dataset covers 21,376 articles with the highest-level section marks as ground-truth segment boundaries. We follow the setting in Arnold et al. (2019) by splitting the dataset into 70% training, 10% validation and 20% test data.

Dataset	# of doc	# sent/seg	# seg/doc
WIKI-50	50	13.6	3.5
Cities	100	5.2	12.2
Elements	118	3.3	6.8
Clinical	227	28.0	5.0

Table 3: Statistics of the datasets used in domain transfer experiments.

4.1.2 Domain Transfer Datasets

To better evaluate models' robustness in cases where a domain-shift is present (called "domain transfer segment inference"), we apply the topic segmenters trained on Wiki-Section to four small corpora heavily deviating from the training corpus (see Table 3 for more details):

Wiki-50 (Koshorek et al., 2018) consists of 50 Wikipedia articles randomly sampled from the latest English Wikipedia dump. There is no overlap between this dataset and Wiki-Section.

Cities (Chen et al., 2009) consists of 100 Wikipedia articles about cities. There is no overlap between this dataset and Wiki-Section, even the theme of this dataset is close to the portion of city articles in Wiki-Section.

Elements (Chen et al., 2009) consists of 118 Wikipedia articles on chemical elements.

Clinical (Malioutov and Barzilay, 2006) consists of 227 chapters in a clinical book. The subsection marks within each chapter are deemed as ground-truth segment boundaries.

4.2 Experimental Design

Baselines: We directly compare our proposed discourse-aware topic segmentation model (called **Basic Model + Discourse**) with the following unsupervised and supervised baselines:

- **BayesSeg** (Eisenstein and Barzilay, 2008): This unsupervised method makes segmentation prediction by situating the lexical cohesion of text in a Bayesian framework. A text span produced by a distinct lexical distribution is recognized as a coherent topic segment.
- **GraphSeg** (Glavaš et al., 2016): This unsupervised method derives semantically coherent segments through reasoning on a semantic relatedness graph construed from greedy lemma alignment.
- **TextSeg** (Koshorek et al., 2018): This supervised neural topic segmenter adopts a hierarchical neural

[7] https://www.govinfo.gov/

sequence labeling framework with BiLSTM as the main architecture of each layer. The basic model used in our paper (described in section 3.1) is an effective extension of this approach.

- **Sector** (Arnold et al., 2019): This is a supervised neural topic segmenter extended from *TextSeg* by adding an auxiliary layer for sentence topic label prediction. The learned intermediate topic embeddings for sentences are directly utilized for segment boundary inference.
- **Transformer** (Glavas and Somasundaran, 2020): This is a supervised neural topic segmenter consisting of two hierarchically connected Transformer networks for sentence encoding and sentence contextualization respectively.
- **Basic Model + Context** (Xing et al., 2020): This is a top-performing neural topic segmenter which shares the same basic architecture with our proposal. The approach improves the **context modeling** capacity of the plain basic model by adding an auxiliary coherence prediction module and restricted self-attention.

Evaluation Metrics: We use the P_k error score[8] (Beeferman et al., 1999) for our intra-domain and domain transfer segment inference evaluations. The metric thereby simply measures the probability that a pair of sentences located at two ends of a k-sized sliding window in a document are incorrectly identified as belonging to the same segment or not. k is determined as half of the average true segment size of the document. Since it is a penalty metric, lower values indicates better performance. We further quantitatively analyze models' efficiency according to two aspects: Model size and model speed, evaluating the count of learnable parameters and batches/documents processed per second during training/inference, besides P_k measurement.

Implementation Details: For the hierarchical topic segmenter (our basic model), we adopt the default setting in Xing et al. (2020), with GoogleNews word2vec ($d = 300$) as initial word embeddings and the contextualized representation of special token `[CLS]` ($d = 768$) from `bert-base-uncased` as initial sentence embeddings. All BiLSTM layers have the hidden state size = 256. For the discourse graph model component, the number of GAT layers is set to 2 through validation and the number of heads is set to 4 as in (Veličković et al., 2018). The input and output dimensions of each layer = 256. Training uses Adam with $lr = 1e^{-3}$ and batch size = 8. Early stopping is applied within 10 epoches of model training and the boundary prediction threshold τ is tuned over the validation set of each corpus we use for intra-domain model evaluation.

Dataset	Choi	Rules	Section	RSTDT
Random	49.4	50.6	51.3	40.5
BayesSeg	20.8	41.5	39.5	37.5
GraphSeg	6.6	39.3	44.9	58.7
TextSeg	1.0	7.7	12.6	26.9
Sector	–	–	12.7	–
Transformer	4.8	9.6	13.6	–
Basic Model	0.81	7.0	11.3	26.9
+Context	**0.54**	**5.8**	**9.7**	25.4
+Discourse	0.59	6.1	10.2	**24.8**

Table 4: P_k (↓) error score on three corpora for intra-domain experiment. Results in **bold** and underlined indicates the best and second best performance across all comparisons. The row in purple is the results achieved by our proposal. The column in green is the results for RSTDT paragraph break prediction with gold discourse structures integrated.

4.3 Intra-Domain Segment Inference

We report our results of the intra-domain segment inference on the Choi, Rules and Wiki-Section datasets in Table 4. For better performance comparison, the table is subdivided into three sub-tables: random baseline, previously proposed approaches and models build on top of the basic model we use. We observe that the basic model without any additinal components already outperforms alternative supervised and unsupervised segmenters. With the above-sentence discourse dependency information injected, as proposed in this paper, the method (named +Discourse) further improves the performance by a notable margin across all three corpora. We further find that our proposed approach does not achieve superior performances compared to the basic model enhanced with the context modeling strategy (+Context) in Xing et al. (2020). We believe that a possible explanation for this underperformance could be the upstream parsing error of the discourse dependency parser applied out-of-domain, oftentimes severly impairing the parsing performance (Huber and Carenini, 2019). Therefore, we conduct an additional experiment on RST-

[8] We also considered *windiff* (Pevzner and Hearst, 2002) as another evaluation metric. Since it was highly correlated with P_k, we omit it and only present performance by P_k to better compare with results reported in previous works.

Dataset	Wiki-50	Cities	Elements	Clinical
Random	52.7	47.1	50.1	44.1
BayesSeg	49.2	36.2	**35.6**	57.2
GraphSeg	63.6	40.0	49.1	64.6
TextSeg	28.5	19.8	43.9	36.6
Sector	28.6	33.4	42.8	36.9
Transformer	29.3	20.2	45.2	35.6
Basic Model	28.7	17.9	43.5	33.8
+Context	**26.8**	**16.1**	39.4	**30.5**
+Discourse	**26.8**	16.9	41.1	31.8

Table 5: P_k (↓) error score on four test corpora for domain transfer experiment. Results in **bold** and underlined indicates the best and second best performance across all comparisons. The row highlighted in purple is the results achieved by our proposal.

DT due to the availability of gold discourse structures annotated by human for this corpus. With no human-annotated topic segment boundaries at hand, we use paragraph breaks contained in RST-DT articles as the ground-truth for training and testing of topic segmentation models. Our results in Table 4 show that the quality of discourse structure is positively correlated with enlarged improvements achieved by our proposal. In this case, the upper bound achieved by integrating gold discourse structures can even outperform the basic model enhanced by context modeling (+Context).

4.4 Domain Transfer Segment Inference

Table 5 presents the performance of simple baselines, previously proposed models and our new approach on the domain transfer task. Similar to the intra-domain segment inference, the Basic Model+Context approach still achieves the best performance across all testing domains except Elements, in which the unsupervised BayesSeg performs superior. However, our +Discourse strategy still leads to improvement over the basic model, and achieves comparable performance to the best model (+Context) on Wiki-50 and Cities. We believe that it gives evidence that injecting discourse dependency structures has potential to enhance the generality of topic segmentation models.

4.5 Efficiency Analysis

	# Params ↓	T-Speed ↑	I-Speed ↑
Basic Model	4.82M	6.90	35.58
+Context	10.93M	1.49	19.23
+Discourse	**7.97M**	**5.44**	**32.85**

Table 6: The efficiency comparison between our proposal and the method proposed in Xing et al. (2020) on the Wiki-Section corpus. These two models share the same basic segmentation framework. **T-Speed** refers the training speed as number of batches processed per second during training stage. **I-Speed** refers the inference speed as number of documents processed per second during inference stage.

Table 6 compares the efficiency of the top two models, comparing our proposed approach (Basic Model+Discourse) against Basic Model+Context. The experiments for these systems were carried out on a Nvidia Telsa V100 16G GPU card. We observe that our strategy of injecting discourse dependency structures can improve model's performance on intra-domain and domain transfer setting, but with less increase of model size and loss of speed compared to +Context. More specifically, adding our discourse graph modeling component on top of the basic model introduces 65% more learnable parameters while the context modeling components in Xing et al. (2020) cause a 127% parameter increasing. On the other hand, discourse graph modeling slightly slows down the speed of model training and inference by 21% and 7.7% respectively, while making more complex context modeling significantly slows down the speed by 78% and 46%. Together with the previous results about model's effectiveness, we can see that our proposed system would be a better option in practical settings where efficiency is critical.

Additionally, we conduct the same set of experiments for the model with both context modeling module and our proposed discourse structure integration (Basic Model+Context+Discourse). The performance of this model always falls in between +Context and +Discourse individually, but with the worst efficiency measured by model size and speed.

5 Conclusion and Future Work

In this paper, we present a neural topic segmentation model with injection of above-sentence discourse dependency structures inferred from a state-of-the-art discourse dependency parser. Different from previously proposed methods, our segmenter leverages the discourse signal by encoding the topical consistency between sentences from a more global and interpretable point of view. Experiments on multiple settings (intra-domain, domain transfer and efficiency comparison) show that our system achieves comparable performance to one of the current top-performing topic segmenters, with much less model size increase and speed degradation.

In the near future, we plan to investigate the synergy between topic segmentation and discourse parsing more comprehensively, by incorporating the type of inter-sentential rhetorical relations and analyzing whether and how this discourse knowledge can enhance supervised topic segmentation frameworks. In the long run, we intend to explore the possibility for discourse parsing to benefit segment topic labeling, which is another important task usually coupled together with topic segmentation to provide the coarse-grained structural information for documents. Particularly, we believe discourse parsing can potentially enhance the step of key phrase extraction in segment topic labeling due to the significant improvement it brings to the related task of name entity recognition (NER) (Jie and Lu, 2019).

Acknowledgments

We thank the anonymous reviewers and the UBC-NLP group for their insightful comments and suggestions. This research was supported by the Language & Speech Innovation Lab of Cloud BU, Huawei Technologies Co., Ltd.

References

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.

Nicholas Asher. 1993. *Reference to abstract objects in discourse*, volume 50. Springer Science & Business Media.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. 2020. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, Online. Association for Computational Linguistics.

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1):177–210.

Marianne Bertrand, Matilde Bombardini, Raymond Fisman, Bradley Hackinen, and Francesco Trebbi. 2018. Hall of mirrors: Corporate philanthropy and strategic advocacy. Technical report, National Bureau of Economic Research.

Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. 2002. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania.

Harr Chen, S.R.K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379, Boulder, Colorado. Association for Computational Linguistics.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (elus). *arXiv: Learning*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information Systems*, 55(3):529–569.

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.

Vanessa Wei Feng and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–521, Baltimore, Maryland. Association for Computational Linguistics.

Michel Galley, Kathleen R. McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 562–569, Sapporo, Japan. Association for Computational Linguistics.

Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In *Proceedings of* *the Fifth Joint Conference on Lexical and Computational Semantics*, pages 125–130, Berlin, Germany. Association for Computational Linguistics.

Goran Glavas and Swapna Somasundaran. 2020. Two-level transformer and auxiliary coherence modeling for improved text segmentation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 2306–2315.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167, Online. Association for Computational Linguistics.

Grigorii Guz, Patrick Huber, and Giuseppe Carenini. 2020. Unleashing the power of neural discourse parsers - a context and structure aware approach using large scale pretraining. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3794–3805, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, et al. 2010. Hilda: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3).

Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9230–9240, Online. Association for Computational Linguistics.

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2306–2316, Hong Kong, China. Association for Computational Linguistics.

Patrick Huber, Linzi Xing, and Giuseppe Carenini. 2021. Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *The Thirty-sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, pages 10794–10802.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Zhanming Jie and Wei Lu. 2019. Dependency-guided LSTM-CRF for named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3862–3872, Hong Kong, China. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

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.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Association for Computational Linguistics*, 6:63–75.

Kelvin Lo, Yuan Jin, Weicong Tan, Ming Liu, Lan Du, and Wray Buntine. 2021. Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3334–3340, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2012. A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1157–1168, Jeju Island, Korea. Association for Computational Linguistics.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Sydney, Australia. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mandar Mitra, Amit Singhal, and Chris Buckley. 1997. Automatic text summarization by paragraph extraction. In *Intelligent Scalable Text Summarization*.

Panitan Muangkammuen, Sheng Xu, Fumiyo Fukumoto, Kanda Runapongsa Saikaew, and Jiyi Li. 2020. A neural local coherence analysis model for clarity text scoring. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2138–2143, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

HyoJung Oh, Sung Hyon Myaeng, and Myung-Gil Jang. 2007. Semantic passage segmentation based on sentence topics for question answering. *Information Sciences*, 177(18):3696–3717.

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

Martin Riedl and Chris Biemann. 2012. How text segmentation algorithms gain from topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 553–557, Montréal, Canada. Association for Computational Linguistics.

Alessandro Solbiati, Kevin Heffernan, Georgios Damaskinos, Shivani Poddar, Shubham Modi, and Jacques Cali. 2021. Unsupervised topic segmentation of meetings with bert embeddings.

Yiping Song, Lili Mou, R. Yan, Li Yi, Zinan Zhu, X. Hu, and M. Zhang. 2016. Dialogue session segmentation by embedding-enhanced texttiling. In *INTERSPEECH*, page 2706–2710.

Xiaofei Sun, Yuxian Meng, Xiang Ao, Fei Wu, Tianwei Zhang, Jiwei Li, and Chun Fan. 2022. Sentence similarity based on contexts. *Transactions of the Association for Computational Linguistics*, 10:573–588.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Yizhong Wang, Sujian Li, and Houfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–188, Vancouver, Canada. Association for Computational Linguistics.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003a. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Bonnie Webber, Matthew Stone, Aravind Joshi, and Alistair Knott. 2003b. Anaphora and discourse structure. *Computational Linguistics*, 29(4):545–587.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3011–3021, Hong Kong, China. Association for Computational Linguistics.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Linzi Xing, Brad Hackinen, Giuseppe Carenini, and Francesco Trebbi. 2020. Improving context modeling in neural topic segmentation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 626–636, Suzhou, China. Association for Computational Linguistics.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. In *The Thirty-fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 14176–14184.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Nan Yu, Meishan Zhang, and Guohong Fu. 2018. Transition-based neural RST parsing with implicit syntax features. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 559–570, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Hainan Zhang, Yanyan Lan, Liang Pang, Hongshen Chen, Zhuoye Ding, and Dawei Yin. 2020. Modeling topical relevance for multi-turn dialogue generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3737–3743. International Joint Conferences on Artificial Intelligence Organization.

Yifei Zhou and Yansong Feng. 2022. Improve discourse dependency parsing with contextualized representations. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2250–2261, Seattle, United States. Association for Computational Linguistics.