WebDP: Understanding Discourse Structures in Semi-Structured Web Documents

Peilin Liu[1,3], **Hongyu Lin**[1]*, **Meng Liao**[4], **Hao Xiang**[1,3], **Xianpei Han**[1,2]*, **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Data Quality Team, WeChat, Tencent Inc., China
{liupeilin2020,hongyu,xianghao2022,xianpei,sunle}@iscas.ac.cn
maricoliao@tencent.com

Abstract

Web documents have become rich data resources in current era, and understanding their discourse structure will potentially benefit various downstream document processing applications. Unfortunately, current discourse analysis and document intelligence research mostly focus on either discourse structure of *plain text* or *superficial visual structures* in document, which cannot accurately describe discourse structure of highly free-styled and semi-structured web documents. To promote discourse studies on web documents, in this paper we introduced a benchmark – **WebDP**, orienting a new task named **Web** Document **D**iscourse **P**arsing. Specifically, a web document discourse structure representation schema is proposed by extending classical discourse theories and adding special features to well represent discourse characteristics of web documents. Then, a manually annotated web document dataset –WEBDOCS is developed to facilitate the study of this parsing task. We compared current neural models on WEBDOCS and experimental results show that **WebDP** is feasible but also challenging for current models. [1]

1 Introduction

With the rapid development of internet during past decades, *web documents* have become one of the most primary and the biggest data resources in current era. As a result, understanding their *discourse structure* – how different components in a document semantically interact with each other to form a cognitive entirety – will greatly benefit many downstream applications, as previous works have demonstrated the virtue of structure information in other types of documents (Chen et al., 2020; Gevа and Berant, 2018; Xing et al., 2022; Frermann and Klementiev, 2019; Zhang et al., 2020).

The *free-styled*, *semi-structured* nature of web documents gives them characteristics different from traditional forms of documents, providing abundant opportunities and challenges for discourse research. On one hand, web documents exhibit more free-styled discourse organization. For instance, it is common for web documents to encompass ***multiple topics*** (Tsujimoto and Asada, 1990), where different *blocks*[2] within the document are loosely connected by implicit semantic relevance or even describe independent topics (Figure 1). The free-styled nature allows discourse with loose structures, multiple topics and weak coherence, in contrast to the compactly-structured, single-topic and strongly-coherent nature of traditional documents. On the other hand, web documents are semi-structured in respect of their HTML markup language and layout of blocks. The content of web documents tend to be organized under ***multiple hierarchies***, usually reflected by the HTML markup hierarchies (Shinzato and Torisawa, 2004; Yoshida and Nakagawa, 2005). However, the semi-structured information provided by HTML markup and layout structures is merely *superficial*, since it does not consistently align with the *underlying* semantic relations in the discourse structure (Figure 1), to which, analogous phenomenon has also been found in plain texts (van der Vliet and Redeker, 2011). In fact, the *inconsistency* between

*Corresponding authors.
[1]Code is available at: https://github.com/qroam/web-document-discourse-parsing

[2]"Blocks" (Tsujimoto and Asada, 1990), or "physical objects" (Cao et al., 2022; Mao et al., 2003) more formally, are customary concepts used in previous document image processing research. Tsujimoto and Asada (1990) defined "blocks" as "*a set of text lines with the same typeface font and a constant line interval*" within a document. In this paper, we study discourse connections among "blocks" in web documents and use this term as a short name for Elementary Discourse Units in the proposed web document discourse schema. A formal definition of these units is presented in § 3.4.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10235–10258
July 9-14, 2023

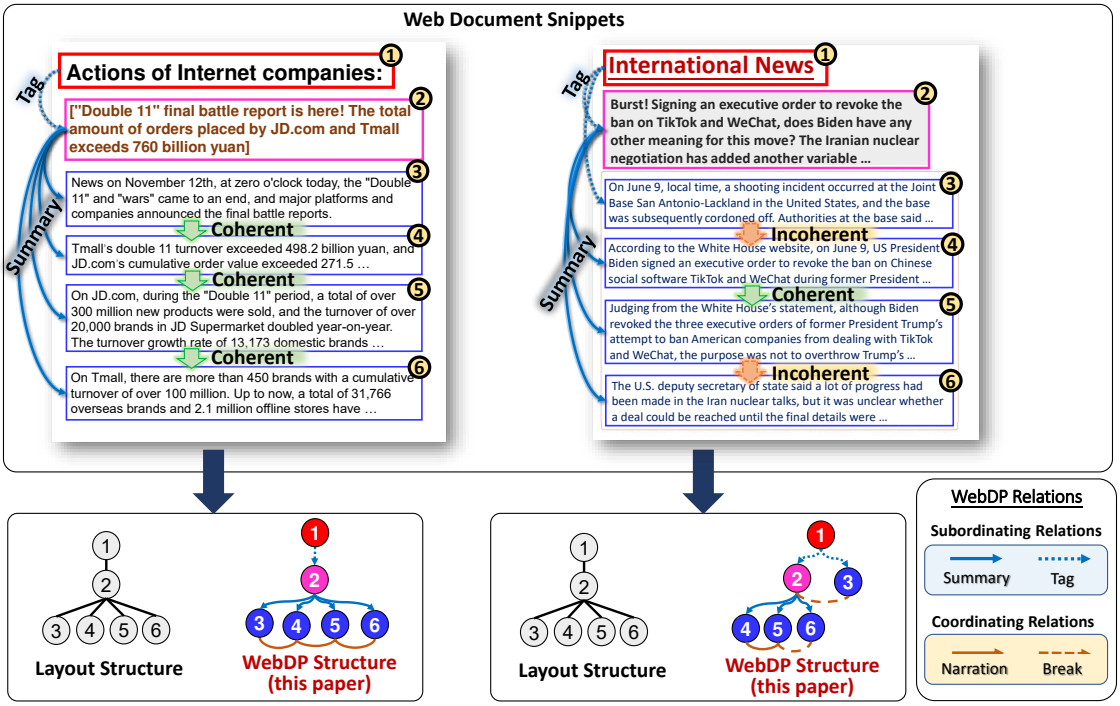

Figure 1: **Characters and challenges in web documents discourse structure representation.** Shown on the top are two representative web document snippets. Web documents, being semi-structured data, naturally incorporate multiple hierarchies of information packaging and exhibit a free-styled nature with multiple topics. These characteristics necessitate the development of a novel discourse schema proposed in this paper (**WebDP Structure**), which effectively captures discourse phenomena in web documents. By contrast, previous studies on superficial visual document structures (e.g., **Layout Structure**) fail to faithfully represent either the discourse relation types (as seen in both documents within the figure) or the discourse structure itself (as observed in the document on the right side). Examples of **WebDP Structures** of real web documents are displayed in appendix F.

these superficial structures and the underlying discourse structure is prevalent in web documents due to their free-styled writing process.

Unfortunately, previous research can not fully fulfill the requirements of discourse representation proposed by the free-styled, semi-structured characteristics of web documents. This limitation can be attributed to two distinct lines of research: Discourse Analysis (Li et al., 2022a) and Document Intelligence (Cui et al., 2021). As for discourse analysis research, traditional studies focus on classical plain texts with compact, single-topic and strong-coherent discourses, leaving a gap to free-styled web documents with loose discourse organization, multiple topics and weak coherence. Furthermore, they mainly study clause-level discourse phenomena, rather than blocks in semi-structured web documents. As for document intelligence research, although they recognize various kinds of semi-structured and multi-modal information carried in documents, including visual, layout, and HTML markup features, most studies only examine superficial structures directly derived from these multi-modal features. As demonstrated in Figure 1, the inconsistency between superficial structures and underlying discourse structure in web documents impairs the power of this line of works in downstream applications that require a deep understanding of semantics.

In this paper, we introduce a new benchmark **WebDP**, based on a well-designed discourse representation schema for web documents named **WebDP structure**. The aim is to promote studies on web document discourse analysis and fill the gap between underlying discourse structure and superficial layout structure in documents. To accurately represent the free-styled discourse structure of web document, we design the discourse schema by extending previous linguistic theories of Rhetorical Relations (Mann and Thompson, 1988) and SDRT (Lascarides and Asher, 2008) to take into account characteristics of discourse organization found in web documents. Specifically, **WebDP structure** adopts two kinds of *rhetorical relations* to connect blocks in document and form a hierarchical structured discourse representation: *subordinating relations* depict semantic functions between blocks across different levels of the hierarchy and

coordinating relations depict coherence properties between blocks at the same-level hierarchy. Then, we collect web documents and manually annotate a dataset WEBDOCS to support further research of **WebDP**. Additionally, to verify the feasibility of the proposed task and reveal its challenges to current methods, we re-implement representative neural models from related fields and conduct systematic experimental analyses on the dataset.

Contribution of this work can be concluded as:

- We formulated a web document discourse schema to explicitly model discourse structures within web documents (§ 3).
- We proposed a new task called **Web** Document **D**iscourse **P**arsing (**WebDP**) based on the web document discourse schema, followed by manually annotating WEBDOCS dataset to support further research (§ 4).
- We conducted systematic experiments and analyses on WEBDOCS to verify the feasibility of the proposed task **WebDP** and reveal its challenges to current methods (§ 5).

2 Related Work

Previous researches approach the structures within documents from two orthogonal research lines, i.e., discourse analysis (Li et al., 2022a) and document intelligence (Cui et al., 2021). Our proposed schema, **WebDP structure**, makes good complement to them on modeling the discourse of web documents (See Table 1 for a brief summary).

2.1 Discourse Analysis

Traditional research on discourse analysis have made comprehensive studies into discourse structure representation based on various kinds of discourse theories (Fu, 2022). Flat (Prasad et al., 2008; Zhou and Xue, 2012; Xue et al., 2015), tree (Carlson et al., 2001; Li et al., 2014; Yoshida et al., 2014; Jiang et al., 2018) and graph (Wolf and Gibson, 2005) structures had been exploited in discourse schemas. Despite their comprehensive discussion on theoretical foundations, traditional studies are mainly confined to *short, single-topic, plain text* and model discourse relations at *clause level*, therefore can not be directly borrowed as solutions to modeling *long, multiple-topic, semi-structured* web documents at *block level*.

There are also researches pay attention to *genre-specific discourse structure*. For example, Dialogue Discourse Parsing (Afantenos et al., 2015; Asher et al., 2016; Li et al., 2020) for modeling discourse structures of multi-turn conversation; discourse structure of technical web forums (Wang et al., 2011), news articles (Choubey et al., 2020) and long-form answers (Xu et al., 2022). In this paper, we take a similar motivation to formulate the discourse structure in web documents by modeling their genre-specific characteristic.

Another line closely related to discourse analysis is Text Segmentation (Choi, 2000; Purver, 2011), which aims to segment long text into shorter segments with inner topic coherence. While text segmentation can be seen as parsing a shallow, linear structure, our proposed schema can provide more abundant semantic representation capability and well account for transition of topic at various hierarchical levels as suggested by Eisenstein (2009).

2.2 Document Intelligence

Document Logical Hierarchy Extraction (or Document Logical Structure Analysis) (Tsujimoto and Asada, 1990; Summers, 1998; Mao et al., 2003; Pembe and Güngör, 2015; Manabe and Tajima, 2015; Rahman and Finin, 2017; Cao et al., 2022) and Table of Content Extraction (Maarouf et al., 2021) aims at parsing document to produce a hierarchical structure based on layout relationship between physical blocks. As we mentioned before, the layout relationship in web documents does not imply consistent semantic relation types. Therefore, the underlying discourse structure of web documents can not be faithfully represented by these task settings.

Wang et al. (2020) proposed Form Understanding which models latent hierarchy in forms by a single "key-value" relation. Similarly, Hwang et al. (2021) models semi-structured document images information extraction as Spatial Dependency Parsing between single tokens. Different from their ideas dedicated to forms and other short document images, our work aims at modeling discourse structure of whole, long document and describes richer categories of semantic relations.

3 Designing Discourse Representation Schema for Web Documents

In this section, we describe the proposed web document discourse schema – **WebDP structure**. We first summarize main characteristics of web documents and the consequent requirements for a well-designed discourse schema (§ 3.1). Then we briefly

Task Name	Text Genre	Basic Unit	Modeling of Semantics	Single Topic Document
Discourse Parsing				
RST (Carlson et al., 2001)	Plain, short text	Clause	✓	*not constrained*
PDTB (Prasad et al., 2008)	Plain, short text	Clause	✓	*yes*
Discourse Dependency Parsing (Yang and Li, 2018)	Plain, short text	Clause	✓	*yes*
Macro-level DP (Jiang et al., 2018)	Plain text	Paragraph	✓	*yes*
Dialogue DP (Asher et al., 2016)	Multi-turn Dialogue	Utterance turn	✓	*yes*
Document Intelligence				
Logical Hierarchy Extraction (Cao et al., 2022)	Long document	Physical Block	✗	*not constrained*
Form Understanding (Wang et al., 2020)	Forms in document	Form Cell	✗	*not constrained*
Spatial Dependency Parsing (Hwang et al., 2021)	Complex-layout documents	Document Token	✗	*not constrained*
Text Segmentation (Purver, 2011)	Long document	Sentence	✗	*not constrained*
WebDP (This Paper)	Long, web document	Physical Block	✓	*not constrained*

Table 1: Comparison of previous task settings which aim at modeling structures within documents and our proposed web document discourse parsing.

review Rhetorical Relations (Mann and Thompson, 1988) and SDRT (Lascarides and Asher, 2008) theories and discuss their merits on modeling web document discourse (§ 3.2). Finally, we extend the above theories by carefully summarizing semantic relation labels special to web documents and propose the new web document discourse schema (§ 3.3 and § 3.4).

3.1 Characteristics of web document discourse structure

As a special genre of text, current web documents have *free-styled discourse organization* and *semi-structured data format*, which bring them unique properties and make traditional task settings no direct solution to their discourse structure. Specifically, we mainly focus on two prominent characteristics caused by the nature of web documents: ***multiple hierarchies*** caused by semi-structured data format and ***multiple topics*** caused by free-styled discourse organization, which are carefully summarized through a preliminary case study on web document instances, and we use them to calibrate a well-designed discourse schema.

Multiple Hierarchies. Documents are intentionally designed to have various levels of information packaging, where some blocks subordinate others semantically or pragmatically to form hierarchical structures. Although such hierarchies of semantics are usually indicated by semi-structured layout and markup features (Power et al., 2003), unfortunately, it is common to have inconsistency between these superficial structures and underlying semantic hierarchy. Therefore, to accurately represent discourse structure of web documents, a schema should *well account for the hierarchical structure to discriminate different underlying semantic functions between blocks in the hierarchy*.

Multiple Topics. It is common for web documents to contain multiple topics of content which are semantically incoherent with each other (Tsujimoto and Asada, 1990). Therefore, besides the hierarchical structure realized by various kinds of semantic functions between different hierarchy levels, a discourse schema for web documents should also *explicitly model the topic transition and semantic incoherence between blocks on the same hierarchy level*.

Furthermore, due to the free-styled and user-generated writing process, web documents could be noise-intensive. To ensure coverage of a wide range of web document instances and robustness against noise, the discourse schema should *strike a balance between expressive capability and universality, necessitating conciseness*.

To meet these requirements, we refer to classical discourse linguistic theories of Rhetorical Relations (RR, Mann and Thompson, 1988; see Jasinskaja and Karagjosova, 2015 for an integrated review) and Segmented Discourse Representation Theory (SDRT, Lascarides and Asher, 2008) to borrow intuition. In the following, we will first briefly review key points of these theories which we grounded to. Then we formulate **WebDP structure** based on them.

3.2 Rhetorical Relation Theory

Mann and Thompson (1988) firstly suggested that *coherence* of text is realized by some *function* that connect each different parts of it. They called these function Rhetorical Relations (RRs). Mann and Thompson (1988) suggested the list of RRs is *po-*

tentially open to addition of new relations, for the sake of describing discourse structure of particular texts. We assume that different blocks in a web document are also connected by some RRs to realize the coherence of document when being read.

Further, two kinds of RRs have been discriminated (Asher and Vieu, 2005), i.e., **Subordinating RRs** and **Coordinating RRs**. 1) **Subordinating RRs**, such as *Elaboration* and *Explanation*, exist between units with *unequal* information packaging levels, where one is subordinate to the other. 2) **Coordinating RRs**, such as *Narration*, *Parallel* and *Contrast*, exist between units of the *same* information packaging level. SDRT theory (Asher, 1993; Asher and Lascarides, 2003; Lascarides and Asher, 2008) adopt the distinction of two types of RRs and consider they govern the hierarchical structure of discourses.

The distinction of subordinating and coordinating RRs provides suitable theoretical foundation for modeling discourse structure of web documents. On one hand, various types of semantic relations between ***multiple hierarchies*** can be modeled by **subordinating RRs** which describe dominating relations between units on unequal information packaging levels. On the other hand, ***multiple topics*** characteristic and different semantic relations between same-level unit can be modeled by **coordinating RRs**. Therefore, we decide to design a web document discourse schema based on RR theory.

3.3 Overview of WebDP Structure

Based on RR theory, we propose the new discourse schema for modeling web document discourse structure. Specifically, **WebDP structure** is composed with *elementary discourse units* of document (i.e., blocks in web document) and binary *rhetorical relations* between them. Two kinds of rhetorical relations are considered in the structure: 1) **Subordinating relations**, which can be analogous to parent-child relation in a tree, and 2) **Coordinating relations**, which can be analogous to relation between successive sibling nodes in a tree. Note that subordinating relations in the proposed discourse structure are not necessarily consistent with the layout structure in web documents.

We can denote the **WebDP structure** as $\mathcal{G}^{(d)} = \{\mathcal{V}^{(d)}, \mathcal{E}^{(d)}\}$, a sparse graph structure consisting node set $\mathcal{V}^{(d)}$ and edge set $\mathcal{E}^{(d)}$. Each edge r_k in $\mathcal{E}^{(d)}$ indicates a binary relation between a pair of nodes, attributed by its rhetorical relation label. Furthermore, edges in $\mathcal{E}^{(d)}$ can be divided into 2 disjoint subsets according to the type of rhetorical relation they hold, i.e. *subordinating edge set* $\mathcal{E}_s^{(d)}$ and *coordinating edge set* $\mathcal{E}_c^{(d)}$.

We further add some constraints on the resulting structures to restrict its complexity. 1) *subordinating edges* are directed since subordinating relations are semantically asymmetric. We define the direction of *subordinating edges* as from subordinated (lower-level in hierarchy) nodes to dominating (upper-level in hierarchy) nodes and stipulate that each node has at most 1 outgoing *subordinating edge* while the number of incoming *subordinating edges* is unconstrained. 2) *coordinating edges* are undirected since coordinating relations are symmetric. We stipulate that each node can be linked by at most 2 *coordinating edges* and there are no cycle formulated by *coordinating edges* themselves.

In the following, we instantiate the schema by specifying the constitution of node set and rhetorical relations in two kinds of edge subsets.

3.4 Elements in WebDP Structure

Nodes in Discourse Schema. Nodes are elementary discourse units (EDUs) in discourse structure when represented as a graph. While traditional discourse parsing usually take clauses as elementary units, we follow document logical hierarchy extraction studies (Cao et al., 2022; Tsujimoto and Asada, 1990; Mao et al., 2003) to model semantic relations between *physical objects* or *blocks* in document. Tsujimoto and Asada (1990) defined *blocks* as "*a set of text lines with the same typeface font and a constant line interval*" within a document. Summers (1998) suggested that segments of document in logical structure should be "*visually distinguished semantic component*" in the document, emphasizing both layout requirement and content requirement.

Following these previous concepts, we formalize "blocks" in **WebDP structure** as non-overlapping physical objects (a region with a bounding box) in documents with two criterion: 1) **Layout homogeneity**, requiring blocks should be visually distinguishable and have clear layout boundaries segmenting them from other parts of the document. 2) **Semantic Coherence**, requiring blocks should have internal semantic coherence in their content. In practice, "blocks" in web document data are usually paragraph-level units. For example, heading, paragraph, items in bullet lists, caption of figures

are several common kinds. Blocks serve as EDUs and constitute the node set $\mathcal{V}^{(d)}$ in our web document discourse schema.

Subordinating Relation Set. We define 7 types of subordinating relations, namely ELABORATION, EXPLANATION, TOPIC&TAG, ATTRIBUTE, LITERATURE, CAPTION and PLACEHOLDER. While the first four relation types also appear in previous discourse schema, we include three additional subordinating relations based on data-driven analysis. These additions aim to address specific cases unique to web documents that cannot be adequately represented by the previously defined relations. Also to be noticed is that in our schema, the subordinating EDUs are blocks with more concise and succinct information, occupying higher information packaging level, which is slightly different from the definition of "nucleus" (EDUs with more significant information) in RST-styled discourse schema. Detailed interpretations for each type of relation can be found in appendix A.1.

Coordinating Relation Set. We define 5 types of coordinating relations, namely NARRATION, LIST, PARALLEL&CONTRAST, TOPIC_CORRELATION and BREAK. These coordinating relations are designed to model various degrees of coherence within long documents between EDUs inside the same information packaging level, forming a spectrum from tight to loose semantic coherence. Among them, LIST is a common discourse expression pattern in documents compared with plain texts; and BREAK is of high frequency in free-styled web documents. Detailed interpretations for each type of relation can be found in appendix A.2.

4 WebDP: A new benchmark for Web Document Discourse Parsing

To facilitate web document discourse analysis research, we present a new task named **Web** document **D**iscourse **P**arsing (**WebDP**). The goal of the task is to automatically convert linear list of EDUs in the input web document into the hierarchical discourse structure defined in **WebDP structure**. To benchmark the task, we construct a new dataset WEBDOCS and introduce its evaluation metrics. In virtue of WEBDOCS, we can benchmark further studies on **WebDP**, reveal the key challenges and difficulties on web document discourse analysis, assess the effectiveness and diagnose the defect of different parsing algorithms.

4.1 Data Collection

A well-designed dataset should widely cover main characteristics of web documents we analysed before. With this consideration, we choose WeChat Official Account[3] as data source based on the following reasons. First, web documents on WeChat Official Account are highly free-styled as they are generated by a multitude of individual authors who contribute content independently, rather than adhering to a standardized editing norm. Specifically, we find these web documents have salient ***multiple topics*** and ***multiple hierarchies*** phenomena and are also enriched with inconsistency between superficial visual structure and underlying discourse structure. Second, WeChat Official Account has a broad and active user community. This makes a dataset based on it universal in domain, extensible in scale and have practical values.

To get the content of each EDU in web documents, HTML source codes are crawled from web sites and we use a naive rule-based parsing script to extract textual content along with their XPath information for each HTML element.

4.2 Data Annotation

In the annotation stage, two human annotators are employed to give golden WebDP discourse structure annotations to web documents. Both annotators are native Chinese (the same language of the data) speakers and hold bachelor's degrees in education. We recruited them through advertising on an internal institution forum. After that, according to the number of applicants, we set a target of 300 web documents in total where each annotator assigned 180 web documents, leaving 60 randomly sampled documents to be doubly-annotated to measure annotation consistency (*Human Baseline* in Table 2).

In the formal annotation phase, annotators are asked to label WebDP structures for web documents following 3 steps: 1) **Annotate EDUs**. To annotate EDUs by identifying blocks of text and aligning their content with XPath information[4]. All EDUs within a web document were organized in a list $\mathcal{V}^{(d)} = \{e_1, ...e_{|d|}\}$ and arranged in correct reading order. 2) **Annotate subordinating relations.** To iterate over $\mathcal{V}^{(d)}$ and consider each EDU.

[3] https://walkthechat.com/wechat-official-account-simple-guide

[4] This step is necessitated by the fact that HTML element automatically parsed during data collection may not always well align with physical blocks in rendered web pages.

For the EDU e_i being considered, annotators assigned it the most appropriate subordinating EDU before it ($e_{j<i}$) based on the available subordinating relation set. If a suitable subordinating EDU could not be found, a dummy node was marked. 3) **Annotate coordinating relations.** For the EDU e_i being considered, annotators identified other EDUs which shared the same subordinating EDU with e_i and before e_i. From these EDUs, they selected the EDU that was most semantically coherent with e_i (usually the nearest one); and then chose the coordinating relation between them according to the coordinating relation set.

In the experiment section below, we split 300 annotated documents into 200/50/50 to serve as train/dev/test sets, respectively. See appendix B for details of data annotation and statistics.

4.3 Evaluation Metrics

Similar to previous dependency parsing tasks (Nivre and Fang, 2017), we apply UAS and LAS for evaluating **WebDP**. These metrics calculate the percentage of correct predicted edges with respect to all predicted edges.

- **UAS** (**Unlabeled Attachment Score**) considers a predicted edge to be correct as long as its two terminal nodes are correct, without concerning relation label of edge.
- **LAS** (**Labeled Attachment Score**) considers a predicted edge to be correct only if both its connecting nodes and relation label are correct. Thus UAS is an upper-bound of LAS.

Notice that **WebDP structure** includes two different sets of edges, i.e., *subordinating edge set* and *coordinating edge set*. Thus, we calculate subordinating, coordinating and overall UAS/LAS for different edge sets respectively. The overall UAS/LAS are micro-averages of subordinating and coordinating UAS/LAS.

5 Experiments

5.1 Task Formulation

One straightforward way to address **WebDP** is to model it with a two-stage pipeline, whereby we firstly predict all subordinating relations within a document to establish a backbone of the discourse structure; then we predict all coordinating relations based on the subordinating structure established in the first stage. In the first stage, *subordinating relation prediction* can be formulated as a conventional discourse dependency parsing task, for which various existing baseline models can be employed. In the second stage, *coordinating relation prediction* can be modeled as a classification task, where we introduce an additional classifier layer. We simply feed representations of node pairs, which are consist of successive sibling nodes on the subordinating structure parsed in the first stage, into this classifier layer to predict their coordinating relation labels. During training, all components of the model can be learned jointly and the classifier in the second stage is trained using the golden subordinating structures as input.

5.2 Baseline Models

For the dependency parsing model in the first stage, we choose baselines from the state-of-the-art models of Dialogue Discourse Parsing (Afantenos et al., 2015) and Document Logical Hierarchy Extraction. To take advantage of the semi-structured HTML markup information contained in web documents, we use **XPath encoders** (Li et al., 2022b; Lin et al., 2020; Zhou et al., 2021) to enrich node representation beyond text encoder.

Specifically, the baseline dependency parsers we choose to compare include: 1) **NodeBased**. A naive solution to the task, simply based on an EDU representation module (*node encoder*) and a node-pair interaction module (*classifier*). 2) **DeepSeq** (Shi and Huang, 2019) designs an *incremental* predicting method to leverage global history information. 3) **Put-or-Skip** (Cao et al., 2022) also adopts incremental decoding while it models the *context information* of each possible insertion site for the current node. 4) **SSAGNN** (Wang et al., 2021) adopts a *fully-connected graph neural network* to enhance the modeling of deep interactions between nodes representation. 5) **DAMT** (Fan et al., 2022) is based on **SSAGNN** and model the dependency parsing task in a *multi-task learning* manner. Details of implementation can be found in appendix C.

5.3 Main Result

We show main experimental results in Table 2. We can see that:

1) **WebDP is a feasible task whose patterns can be learned by neural methods.** Current methods have an UAS of 60-65 and LAS of 50-55, which are moderate values refer to other document-level discourse parsing tasks (Li et al., 2014; Afantenos et al., 2015) and indicate the feasibility of **WebDP**. Thus, the discourse structure we defined on web documents can be effectively learned and

models trained on **WebDP** have potential to be utilized on downstream tasks.

2) **WebDP is a challenging task and current methods leave great room for improvement.** Compared with human baseline in Table 2, state of the art parsers still lag far behind, leaving great room for improvement on **WebDP**. We believe this may because **WebDP** models discourse structure unique to the free-style genre of web documents and considers discourse relations at a more macroscopic block-level. In the future, designing task-specific models for better modeling unique features of web documents could be worthy of study.

5.4 Detailed Analysis

To further investigate what kinds of documents or EDU instances are more challenging to current models, we conduct analysis from three aspect: *the influence of document length*, *the influence of dependency edge spanning distance* and the *influence of multiple topics* on **WebDP** performance. From results plotted in Figure 2, we can see that:

1) **Long web documents pose challenge to current models.** As Figure 2 (a) shows, with the increasing of document length, instance level performance drops gradually in general for all the models. Compared with previous discourse tasks, web documents are usually longer and contain numerous nodes. Specifically, previous Discourse Dependency Parsing datasets have an average 15-20 EDUs per document (Nishida and Matsumoto, 2022); and for Dialogue Dependency Parsing datasets, it is less than 10 (Li et al., 2020). However, in WEBDOCS corpus, we have an average number of 47 EDUs per document some can be even longer than 100 EDUs (Table 6). The larger size and more abundant information in single web document make this task different from its previous analogues and challenging to previous models.

2) **Models suffer from problem of poorly modeling long term dependencies.** Previous work on Dialogue Dependency Parsing (Fan et al., 2022) demonstrated that EDU-level parsing performance has a negative correlation with the distance of golden dependency edge. Here, we also found in **WebDP** such a strong negative correlation as Figure 2 (b) shows. Long term dependencies are hard to learn due to the distance bias introduced by data – – EDUs that are close in space are more frequently linked together. How to effectively remedy the long term dependencies problem needs further studies.

3) **Multiple topics phenomena may come together with more straightforward hierarchy structures, and thus easier for current models.** To understand the influence of ***multiple topics*** on parsing performance, in Figure 2 (c) we plot document-level performance with respect to the proportion of BREAK labels in coordinating edge set (on behalf of the topic change frequency). Figure 2 (c) shows an roughly positive correlation. This might because that documents aggregated with ***multiple topics*** tend to have more explicit and concise hierarchy structure patterns, and the semantic discrepancy between incoherent topics is also obvious enough and easy to capture. On the other hand, how to correctly model ***multiple hierarchies*** within single coherent topic seems more challenging to current models.

5.5 Error Analysis

Models predict wrong structure occasionally, however, errors may not equal with each other. We define 5 kinds of structure error types to depict features of different *subordinating edge prediction errors* in Table 3. Structure error types are defined based on the predicted edges with respect to *golden* structure. From Table 3 we can conclude that:

1) **Models tend to predict dummy subordinating nodes.** Among wrongly predicted subordinating edges, a huge proportion is because models consider that there is no subordinating node for the target node (*Dummy*, 25.89%). On one hand, this may caused by the class imbalance problem in data – due to ***multiple topics***, many nodes in the dataset have no subordinating node (Table 6). On the other hand, compared with previous discourse parsing task where *explicit relations* are abundant, there may be more *implicit relations* (Dai and Huang, 2018) between blocks in documents, which are more difficult for current models to capture.

2) **Models learn semantic correlation and superficial structures from data, while precise information-packing hierarchies underneath semantic correlation is hard to master.** Table 3 indicates that although models wrongly select subordinating nodes, there are around half of the wrongly selected subordinating nodes are closely related to target node *either in semantics or in layout structure*. For example, model tend to consider coordinating nodes as subordinating node (*Sibling*, 22.81%); mistake coordinating nodes of the golden

Compared Models		Subordinating		Coordinating		Overall UAS	Overall LAS
		UAS	LAS	UAS	LAS		
NodeBased		63.18	54.35	63.20	56.07	63.19	55.21
DeepSeq (+SRE)	(Shi and Huang, 2019)	62.35	53.12	63.12	56.06	62.73	54.59
Put-or-Skip	(Cao et al., 2022)	59.79	51.93	65.48	57.04	62.63	54.49
SSAGNN (+GNN)	(Wang et al., 2021)	61.97	53.89	63.95	55.81	62.96	54.85
DAMT (+GNN+multi-task)	(Fan et al., 2022)	61.63	52.90	63.86	54.19	62.74	53.54
Human Baseline		**88.98**	**83.88**	**88.49**	**78.95**	**88.74**	**81.42**

Table 2: **Main Results.** All compared models are equipped with an additional coordinating relation classifier described in § 5.1 and all reported results for compared models are averaged from 3 random seeds. Human baseline is calculated from doubly-annotated document instances. Best performances among compared models are underlined.

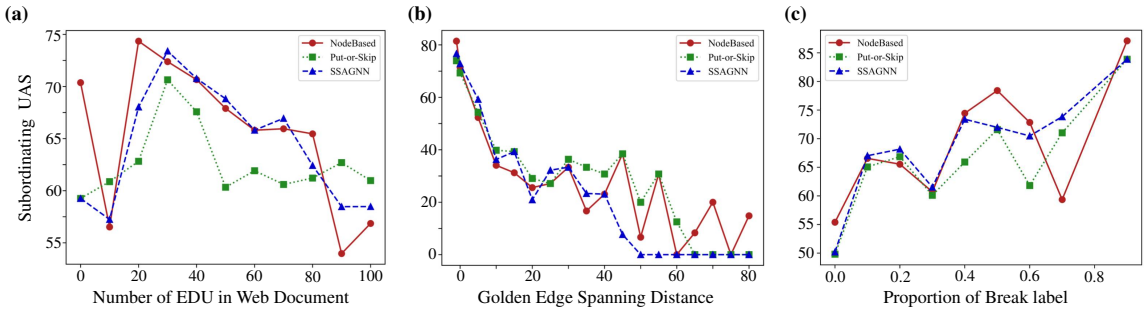

Figure 2: Detailed Analysis on the challenges of **WebDP**. (a) Document-level Subordinating UAS with respect to number of EDUs in document. (b) EDU-level Subordinating UAS with respect to the spanning distance of golden edge. (c) Document-level Subordinating UAS with respect to the frequency of independent topics (proportion of *Break* label). Metric plotted here is the **Subordinating UAS**, see appendix D for same analysis on other metrics.

Error Type (Proportion (%))	Example
Dummy (25.89)	**T**: Contact number: 1********** **GS**: Welcome to inquire: **PS**: (*Dummy Node*)
Sibling (22.81)	**T**: 4. test sites arrangement in area 04 **GS**: The arrangement of test sites in different areas is as follows: **PS**: 3. test sites arrangement in area 03
Ancestor (20.33)	**T**: GACC: In the first seven months ... In particular, China's export to the United States was 1.56 trillion yuan, down by 4.1%; ... **GS**: GACC: China's exports to the United States fell 4.1 percent in the first seven months of this year **PS**: Finance and Economics
Sibling of Parent (11.13)	**T**: In fact, life does not need to be indomitable. Blogger Austin Kleon discovered that ... **GS**: "I have to be more productive" **PS**: "I have to be a better version of myself"
Others (19.82)	**T**: After May 25, candidates can log on the website http://*** to check their test site information. **GS**: Hefei urban college entrance examination test site announced **PS**: In addition, special test site arrangement

Table 3: **Statistics of Structural Error Types in subordinating relations.** **T**: Target EDU, **GS**: Golden Subordinating EDU, **PS**: Predicted Subordinating EDU. Shown here are output cases of **NodeBased**, different compared models have similar error types profile. Examples have been translated into English to display.

subordinating node, which usually have similar layout or superficial language features (*Sibling of Parent*, 11.13%); or select indirect subordinating nodes across several hierarchy levels (*Ancestor*, 20.33%). Both of those error types can be attributed to some semantic correlation or superficial structure shortcuts learned by models, indicating that they still lack exquisite semantic discrimination ability.

6 Conclusion

In this paper, we inspect web documents from a discourse linguistic perspective to reveal their underlying discourse structure. Inspired by linguistic theories of Rhetorical Relations and SDRT, we build a web document discourse schema which simultaneously models subordinating relations and coordinating relations. Based on the schema we propose a new task **WebDP** and contribute a dataset to promote the discourse analysis researches on web documents. Experimental results of recent neural models exhibits the challenge of **WebDP** and detailed analyses provide insights for future studies. We believe the web document discourse schema is prospective in facilitating document-level natural language processing research by explicitly modeling discourse structure for web documents.

Limitations

There may be some possible limitations in this study:

1. Discourse Schema. It should be acknowledged that web documents are heterogeneous themselves and a unified framework to accommodate all web documents may be infeasible. In this article, instead of pre-define the domain/type of web documents we target at, we adopt a problem-motivated research paradigm where we ground ourselves to two characteristics (multiple topics and multiple hierarchies) during discourse schema design. Although we simplify the discourse schema to promote its universality, due to the free-style and domain diversity of web document data, it still has a limited scope of usage, mainly on general news report with multiple topics. For future studies, label sets could be revised in order to better account for the semantic functions in web documents of specific domains, where fine-grained labels and domain-specific labels can be considered.
2. Task Setting. In this paper, we only consider parsing the discourse from a list of document logical blocks already pre-processed in advance while do not contain a complete pipeline from input HTML source code to the final output discourse structure in the task setting. In the future, the gap between HTML elements and document logical blocks should be automatically closed in order to apply to downstream application scenarios.
3. Data Bottleneck. The annotated data volume in this paper is not big enough due to the expensive labour overhead, which may introduce noise into experiments and distort the performance and analysis. Also, the domain diversity and multilingualism of dataset could be questioned since we collect data from single platform in Chinese. In the future, such data bottleneck can be remedy by more dedicated manual annotation efforts, the help of weak supervision techniques, as well as developing data-efficient models.

Ethics Statement

In consideration of ethical concerns, we provide the following detailed description:

1. All of the collected web documents come from publicly available sources. Our legal advisor and/or the web platform confirms that our data source are freely accessible online without copyright constraint to academic use. The data collection protocol has examined and approved by ethics review board.
2. WEBDOCS contains 300 annotated documents. We have done double-checking to guarantee that WEBDOCS contains no sample that may cause ethic issues or involve any personal sensitive information. We also manually check the content of each document instance to exclude any hate speech or attack on vulnerable groups.
3. We hired 2 annotators who have bachelor degrees. Before formal annotation, annotators were asked to annotate 20 document instances randomly extracted from the dataset, and based on average annotation time we set a fair salary (i.e., 35 dollars per hour, which is adequate given the their demographic) for them. During the annotation training process, they were paid as well.

Acknowledgements

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDA27020200 and the National Natural Science Foundation of China under Grants no. 62122077, 62106251.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multiparty chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

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

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Nicholas Asher and Laure Vieu. 2005. Subordinating and coordinating discourse relations. *Lingua*,

115(4):591–610. Coordination: Syntax, Semantics and Pragmatics.

Rong-Yu Cao, Yi-Xuan Cao, Gan-Bin Zhou, and Ping Luo. 2022. Extracting variable-depth logical document hierarchy from long documents: Method, evaluation, and application. *Journal of Computer Science and Technology*, 37(3):699–718.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. 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*.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Lei Cui, Yiheng Xu, Tengchao Lv, and Furu Wei. 2021. Document AI: benchmarks, models and applications. *CoRR*, abs/2111.08609.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Jacob Eisenstein. 2009. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 353–361, Boulder, Colorado. Association for Computational Linguistics.

Yaxin Fan, Peifeng Li, Fang Kong, and Qiaoming Zhu. 2022. A distance-aware multi-task framework for conversational discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 912–921, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Yingxue Fu. 2022. Towards unification of discourse annotation frameworks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 132–142, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mor Geva and Jonathan Berant. 2018. Learning to search in long documents using document structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 161–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-tasking dialogue comprehension with discourse parsing. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 551–561, Shanghai, China. Association for Computational Lingustics.

Wonseok Hwang, Jinyeong Yim, Seunghyun Park, Sohee Yang, and Minjoon Seo. 2021. Spatial dependency parsing for semi-structured document information extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 330–343, Online. Association for Computational Linguistics.

Katja Jasinskaja and Elena Karagjosova. 2015. Rhetorical relations. *The companion to semantics. Oxford: Wiley*.

Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. 2018. MCDTB: A macro-level Chinese discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3493–3504, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022a. A survey of discourse parsing. *Front. Comput. Sci.*, 16(5).

Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. 2022b. MarkupLM: Pre-training of text and markup language for visually rich document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6078–6087, Dublin, Ireland. Association for Computational Linguistics.

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25–35, Baltimore, Maryland. Association for Computational Linguistics.

Bill Yuchen Lin, Ying Sheng, Nguyen Vo, and Sandeep Tata. 2020. Freedom: A transferable neural architecture for structured information extraction on web documents. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 1092–1102, New York, NY, USA. Association for Computing Machinery.

Ismail El Maarouf, Juyeon Kang, Abderrahim Ait Azzi, Sandra Bellato, Mei Gan, and Mahmoud El-Haj. 2021. The financial document structure extraction shared task (FinTOC2021). In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pages 111–119, Lancaster, United Kingdom. Association for Computational Linguistics.

Tomohiro Manabe and Keishi Tajima. 2015. Extracting logical hierarchical structure of html documents based on headings. *Proc. VLDB Endow.*, 8(12):1606–1617.

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.

Song Mao, Azriel Rosenfeld, and Tapas Kanungo. 2003. Document structure analysis algorithms: a literature survey. In *IS&T/SPIE Electronic Imaging*.

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.

Joakim Nivre and Chiao-Ting Fang. 2017. Universal Dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 86–95, Gothenburg, Sweden. Association for Computational Linguistics.

F Canan Pembe and Tunga Güngör. 2015. A tree-based learning approach for document structure analysis and its application to web search. *Natural Language Engineering*, 21(4):569–605.

Richard Power, Donia Scott, and Nadjet Bouayad-Agha. 2003. Document structure. *Computational Linguistics*, 29(2):211–260.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Matthew Purver. 2011. Topic segmentation. *Spoken language understanding: systems for extracting semantic information from speech*, pages 291–317.

Muhammad Mahbubur Rahman and Tim Finin. 2017. Understanding the logical and semantic structure of large documents. *CoRR*, abs/1709.00770.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Keiji Shinzato and Kentaro Torisawa. 2004. Acquiring hyponymy relations from web documents. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 73–80, Boston, Massachusetts, USA. Association for Computational Linguistics.

Kristen Maria Summers. 1998. *Automatic discovery of logical document structure*. Cornell University.

S. Tsujimoto and H. Asada. 1990. Understanding multi-articled documents. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume i, pages 551–556 vol.1.

Nynke van der Vliet and Gisela Redeker. 2011. Complex sentences as leaky units in discourse parsing. *Proceedings of Constraints in Discourse*, pages 1–9.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3943–3949. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Li Wang, Marco Lui, Su Nam Kim, Joakim Nivre, and Timothy Baldwin. 2011. Predicting thread discourse structure over technical web forums. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 13–25, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Zilong Wang, Mingjie Zhan, Xuebo Liu, and Ding Liang. 2020. DocStruct: A multimodal method to extract hierarchy structure in document for general form understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 898–908, Online. Association for Computational Linguistics.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Linzi Xing, Patrick Huber, and Giuseppe Carenini. 2022. Improving topic segmentation by injecting discourse dependencies. In *Proceedings of the 3rd Workshop on Computational Approaches to Discourse*, pages 7–18, Gyeongju, Republic of Korea and Online. International Conference on Computational Linguistics.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3556–3572, Dublin, Ireland. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task*, pages 1–16, Beijing, China. Association for Computational Linguistics.

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.

Minoru Yoshida and Hiroshi Nakagawa. 2005. Reformatting web documents via header trees. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 121–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar. Association for Computational Linguistics.

Nan Yu, Guohong Fu, and Min Zhang. 2022. Speaker-aware discourse parsing on multi-party dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5372–5382, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. 2020. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, Online. Association for Computational Linguistics.

Yichao Zhou, Ying Sheng, Nguyen Vo, Nick Edmonds, and Sandeep Tata. 2021. Simplified DOM trees for transferable attribute extraction from the web. *CoRR*, abs/2101.02415.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77, Jeju Island, Korea. Association for Computational Linguistics.

A List of Discourse Relation Labels

A.1 Subordinating Relations

See Table 4 for the proposed subordinating semantic relations, 7 relation types are designed based on preliminary case study on web documents. Some relation types are borrowed from previous theoretical research while others (LITERATURE, CAPTION and PLACEHOLDER) are added in order to fully account for pragmatic phenomena unique to web document domain. Since the directed subordinating edges can be analogue to parent-child relations in a tree, in Table 4 we use the term "parent node" to refer to the nodes at higher information packaging level (dominating/subordinating nodes) and "child node" to refer to the nodes at lower level (subordinated nodes).

Relation Type	Description	Proportion(%)
Elaboration	The child node provides a detailed elaboration of the semantic content expressed in the parent node. It could involve situations where the child node is summarized by the its parent node completely or partially; where multiple incoherent child nodes are semantically aggregated by one parent node; where the child node restates the same or similar text as the parent node.	46.32
Explanation	The child provides explanations to the parent, provides richer and more detailed information supporting the claim of parent node; or answers questions proposed by parent node.	2.74
Topic&Tag	The relationship between nodes is abstract and conceptual. The parent node is usually an entity, concept, or category. The parent node gives a classification tag for the child node or topically give rise to child node. Since semantic information in the parent node is highly abstract, it cannot be considered as a valid summary of the child node.	15.52
Attribute	The child node is the attribute value of the parent, providing the content referred to by the parent node. It could involve situations where the child node is cited by the parent node; where the parent node is title of a specific genre of text (e.g., "notice", "declaration") and the child node provides corresponding content; where parent node and child node form a key-value pair relationship in a table.	26.22
Literature	This category includes titles commonly found in literary special reports and literary works. The semantics of these parent nodes are primarily literary in nature. They can be considered as placeholders with literary significance to catch attention to the content expressed by their child nodes. Due to their unique nature, they cannot be categorized into any of the above categories.	3.66
Caption	The child node provides caption texts which descriptive information or meta information for the parent node, common in caption texts below images.	4.45
Placeholder	The parent node does not carry actual semantics; however, its existence allows the child nodes to be integrated into a semantic whole. In this relation, there should be a coherent semantic relationship among the child nodes, and there should be clear semantic boundaries between the group of child nodes and other nodes at the same level.	1.09

Table 4: List of subordinating rhetorical relations between different-level nodes.

A.2 Coordinating Relations

See Table 5 for the proposed coordinating semantic relations, 5 relation types standing for various degrees of semantic coherence are designed based on preliminary case study on web documents. Particularly, the label of BREAK, which stands for totally incoherence situations, is added for discriminating ***multiple topics*** phenomena from other coherence discourses which we found widespread in web documents.

B Dataset Annotation Details

B.1 Data Collection and Annotation

We collect web documents within a time window from Jun. 2021 to Oct. 2022 from WeChat Official Account and filter them to discard documents with excessive noise in semantic. The language of collected web documents are mainly in Chinese and the domian is unrestricted. The resulting dataset contains 300 web documents.

HTML source codes are crawled from web sites and we use a naive rule-based parsing script followed by simple manual post-processing to extract content of each HTML element from HTML files. Thanks to the structured property of HTML, by this mean we can easily acquire the content of logical blocks in correct reading order on with lightweight manual post-processing.

We employ two human annotators who have bachelor degrees to give gold discourse structure annotation to the 300 web documents. The annotation stage is composed of an annotator training phrase and a formal annotation phase. During the training phase, annotators are trained with an annotation guideline and a few representative examples which are picked up during preliminary case study. We discuss these examples with the annotators to clear up confusions and revised the final version of annotation guideline. Then, two annotators started to annotate 180 web documents independently in the formal annotation phase. We set a fair salary (35 dollars per hour) to pay the annotators and the training phase is paid as well.

B.2 Data Statistics

See Table 6 for statistical information of the annotated corpus WEBDOCS corpus. Compared with previous discourse parsing tasks, web documents have far more EDUs and each EDU usually contains much longer content, making **WebDP** a challenging task .

C Experiment Details

All of the 5 compared baselines are re-implemented using PyTorch[5] deep learning framework and Hugging Face[6] for loading of pre-trained language model checkpoints. During re-implementation, we adapt all models to the fine-tuning paradigm that adopt a pre-trained BERT text encoder (Devlin et al., 2019). Experiments are performed on single GPU TITAN RTX of 24GB memory and a processor Intel(R) Xeon(R) CPU E5-2683 v3 @ 2.00GHz.

We trained all baseline models for 100 epochs using a batch size of 1 and a linear learning rate scheduler with warm-up. The optimizer we use is AdamW. We set the max EDU number of documents to 200 for both training and evaluation. During training, performance on dev set are evaluated by each epoch, and we choose the epoch checkpoint of best dev set performance to report the final performance on test set. We also did brief hyper-parameter searching on dev set to find that a learning rate of 1e-4 and gradient accumulate steps of 8 are preferable.

D Supplements to Detailed Analysis

Demonstrated in Figure 3, 4 and 5 are the results on other evaluation metrics for the analysis in § 5.4. Conclusions similar to § 5.4 can be drawn, except for the chaotic pattern of Coordinating UAS and Coordinating LAS on the proportion of break label (Figure 4 (c) and Figure 5 (c)). It might because we use pipeline modeling methods which do not directly learn to predict coordinating edges, thus the tendency on Coordinating UAS/LAS appears with more noise.

E Ablation Study

We further conduct a series of ablation studies to investigate the effect of some common practices in previous parsing tasks, as well as whether the introducing of HTML markup feature information unique to that is unique to web documents can benefit the task. The ablation studies are conducted based on **NodeBased** model, which is simply composed with a *node encoder* and a *classifier module* (Table 7. Line 1 "Basic Settings" indicates the ablation setting we reported in main result Table 2 which has best Overall LAS among all settings).

[5] https://pytorch.org
[6] https://huggingface.co/models

Relation Type	Description	Proportion(%)
Narration	Coordinating nodes are coherent in a narrative structure, where information are put in some logical order. Context, impact, and commentary related to the event are also included in this relation type.	42.20
List	Coordinating nodes are a list of parallel points. As a whole, the items listed are equally important and indispensable in order to form the integrity of the whole content.	7.49
Parallel&Contrast	Coordinating nodes linked by comparability in content, as juxtaposition or contrast of similar category of things.	0.41
Topic_Correlation	Coordinating nodes linked by the relevance of the content, and topics move from one to another by this relevance.	1.78
Break	Coordinating nodes have no coherence at all, so they can exist as two completely independent discourses.	48.11

Table 5: List of coordinating rhetorical relations between same-level nodes.

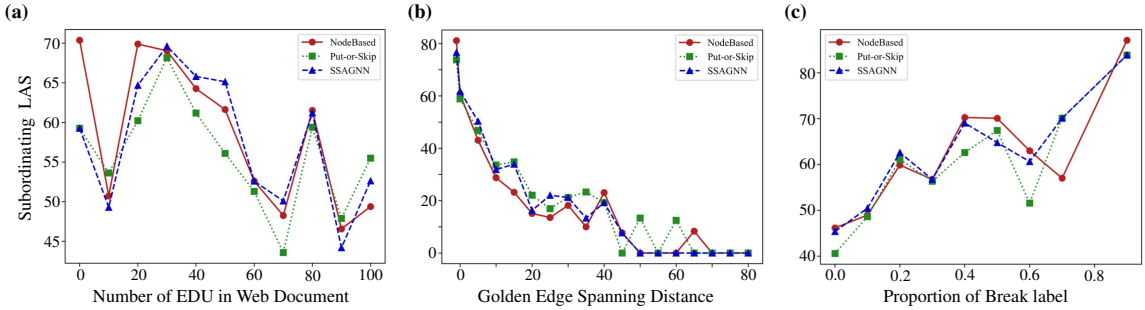

Figure 3: Detailed Analysis on the challenges of **WebDP** similar with Figure 2. Metric plotted here is the **Subordinating LAS**. (a) Document-level Subordinating LAS with respect to number of EDUs in document. (b) EDU-level Subordinating LAS with respect to the spanning distance of golden edge. (c) Document-level Subordinating LAS with respect to the frequency of independent topics (proportion of *Break* label).

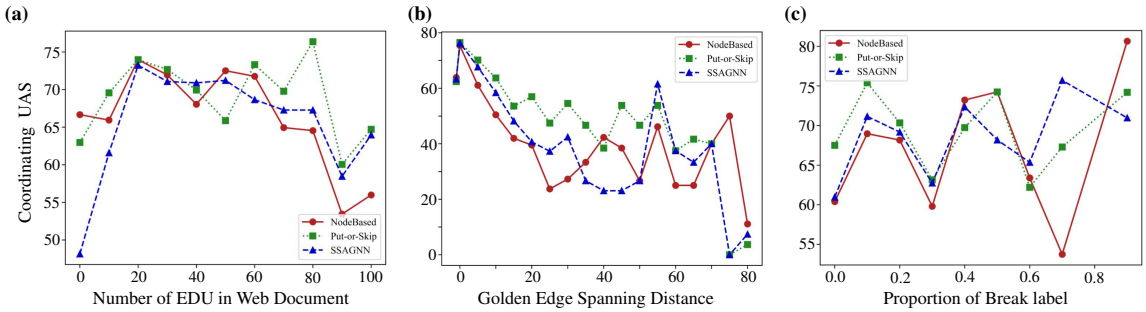

Figure 4: Detailed Analysis on the challenges of **WebDP** similar with Figure 2. Metric plotted here is the **Coordinating UAS**. (a) Document-level Coordinating UAS with respect to number of EDUs in document. (b) EDU-level Coordinating UAS with respect to the spanning distance of golden edge. (c) Document-level Coordinating UAS with respect to the frequency of independent topics (proportion of *Break* label).

As previous researches on web documents have mentioned (Li et al., 2022b; Lin et al., 2020; Zhou et al., 2021), introducing markup information such as HTML tag and XPath is helpful to web documents information extraction, we also believe the markup features may help **WebDP** since the understanding of documents by human intuitively takes the synergy of semantic content and layout manifestation. Here, we adopt XPath encoding methods proposed by these works, i.e., **RNN XPath en-**

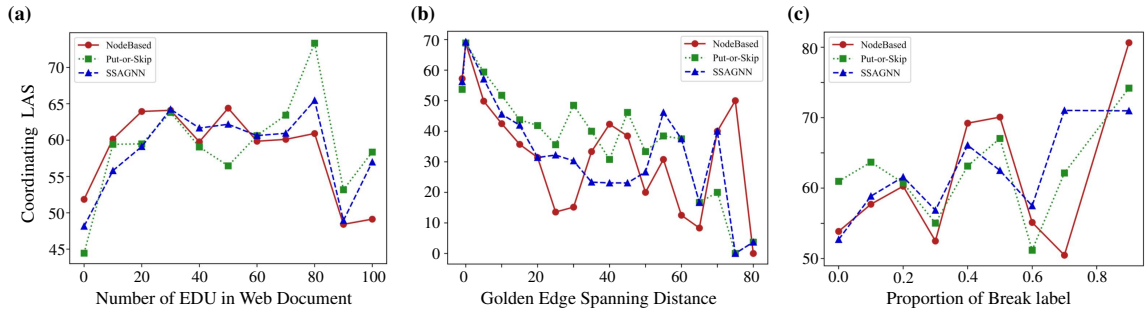

Figure 5: Detailed Analysis on the challenges of **WebDP** similar with Figure 2. Metric plotted here is the **Coordinating LAS**. (a) Document-level Coordinating LAS with respect to number of EDUs in document. (b) EDU-level Coordinating LAS with respect to the spanning distance of golden edge. (c) Document-level Coordinating LAS with respect to the frequency of independent topics (proportion of *Break* label).

Metric	Value
# Document	300
# EDU	14101
Avg/Max. EDU per document	47/174
Avg/Max. document length (in *tokens*)	2204/8756
Avg/Max. EDU length (in *tokens*)	47/823
Avg/Max. dependency distance	6.33/159
# EDU without subordinating node	3493 (24.7%)
Avg/Max. node depth	1.49/8

Table 6: Data Statistics of the WEBDOCS Dataset.

coder (Lin et al., 2020; Zhou et al., 2021) and **FFN XPath encoder** (Li et al., 2022b) to enhance our node representation module. Different modality aggregation methods (Wang et al., 2020; Yu et al., 2022) are also investigated.

For improving the modeling of text information, based on the observation that text piece in nodes often include one or several sentences, we equip models with more advanced sentence representation modules such as **SentenceBERT** (Reimers and Gurevych, 2019) and **SimCSE** (Gao et al., 2021).

Beside adding Xpath information and advanced sentence embeddings to enhance the *node encoder*, we also investigate the effect of **Global Context Encoder** and **Biaffine Attention** mechanism. **Global Context Encoder** (e.g., hierarchical GRU, Wang et al., 2021) is a common practice in discourse-level dependency parsing which add a global interaction layer beyond the representation of each single nodes to model higher level context information. **Biaffine Attention** (Dozat and Manning, 2016) is introduced to replace MLP in *classifier module* and has the merits of effectively and explicitly modeling interaction between node-pairs. Both of them are means of better modeling interaction between nodes either during encoding or predicting.

Results in Table 7 prove our hypotheses in that: 1) **Global Context Encoder** is significantly helpful for current model (Line 1 vs Line 3 in Table 7), since it can effective model the context interaction of EDUs in web documents which are usually too long to be concatenated together and loaded into window size of current PLM encoders as previous works (Line 1 vs Line 5) (He et al., 2021; Fan et al., 2022); 2) using **Biaffine Attention** as classifier is more beneficial than MLP for **WebDP** (Line 1 vs Line 6), which is in accordance with previous observations; 3) different **XPath encoding** and modality aggregation methods all outperform baseline without using XPath (Line 1, 8, 9, 10 vs line 7), indicating the addition of such markup information unique to web documents is helpful to the new task. However, 4) what is unexpected is that pre-trained sentence encoding methods based on modification of BERT do *not* outperform the BERT-base encoder and even damage performance by some margin (Line 1 vs Line 12, 13), this exception may be due to the domain discrepancy between pre-training and downstream data and reason behind need further investigation.

F Examples of Annotated WebDP Structures

Figure 6 and Figure 8 display the WebDP discourse structures of selected web documents in WEBDOCS dataset, and their corresponding English translation versions are presented in Figure 7 and Figure 9.

	Subordinating		Coordinating		Overall UAS	Overall LAS
	UAS	LAS	UAS	LAS		
1. **NodeBased (Basic Settings)**	63.18	54.35	63.20	56.07	63.19	55.21
2. *w/o previous loss*	64.21	53.68	-	-	-	-
Global Encoder						
3. *w/o Global Encoder*	54.93	48.32	54.23	47.65	54.58	47.98
4. *+Transformer Encoder*	54.07	47.18	53.95	47.92	54.01	47.55
5. *+All Concatenated PLM*	49.33	40.45	49.06	43.11	49.20	41.78
Classifier						
6. *+Concatenate MLP*	59.79	51.37	61.55	54.38	60.67	52.88
Multi-modalily Information						
7. *w/o XPath information*	57.71	49.92	59.91	53.29	58.81	51.60
8. *+XPath FFN (Linear Combination)*	61.39	52.11	61.59	53.72	61.49	52.92
9. *+XPath RNN (Concatenation)*	62.61	53.92	62.57	55.64	62.59	54.78
10. *+XPath RNN (Linear Combination)*	63.43	55.36	62.92	54.58	63.18	54.97
11. *XPath only (w/o text information)*	38.88	29.48	48.67	35.43	43.78	32.46
Sentence-level Text Representation						
12. *+SBERT*	56.02	50.23	57.56	48.43	56.79	49.33
13. *+SimCSE*	60.92	51.17	59.98	52.58	60.45	51.88

Table 7: **Ablation Studies.** We demonstrate that the Global Encoder which models document-level global context of EDUs; the Biaffine Classifier which models pairwise interaction at decoding stage; as well as the semi-structured XPath information which is unique to web documents are complementally helpful for **WebDP** task.

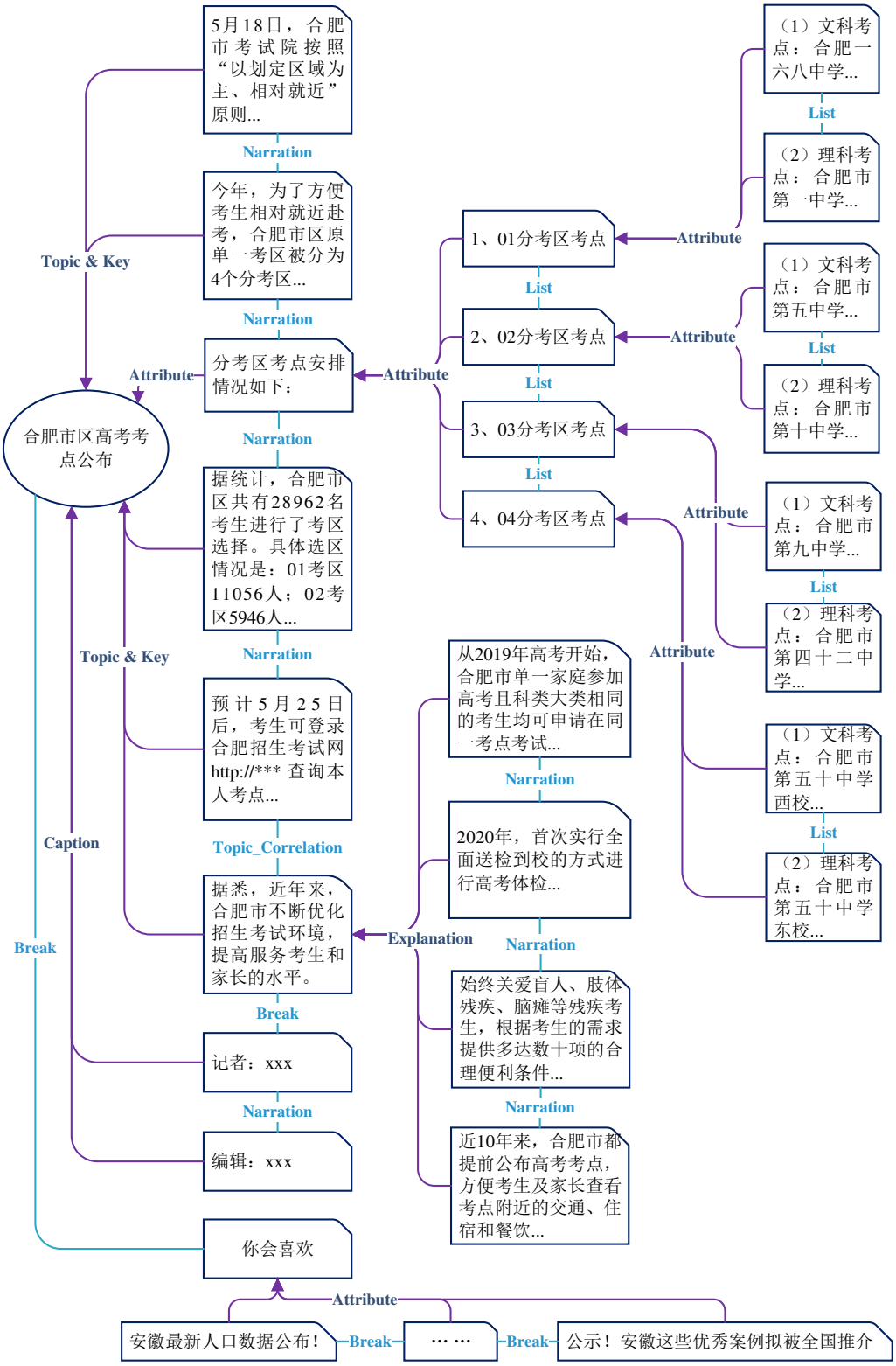

Figure 6: Selected discourse structure example from WEBDOCS dataset. Some content are omitted with ellipsis for clearer display. An English translation version of the same web document discourse structure and web document snippet can be found in Figure 7.

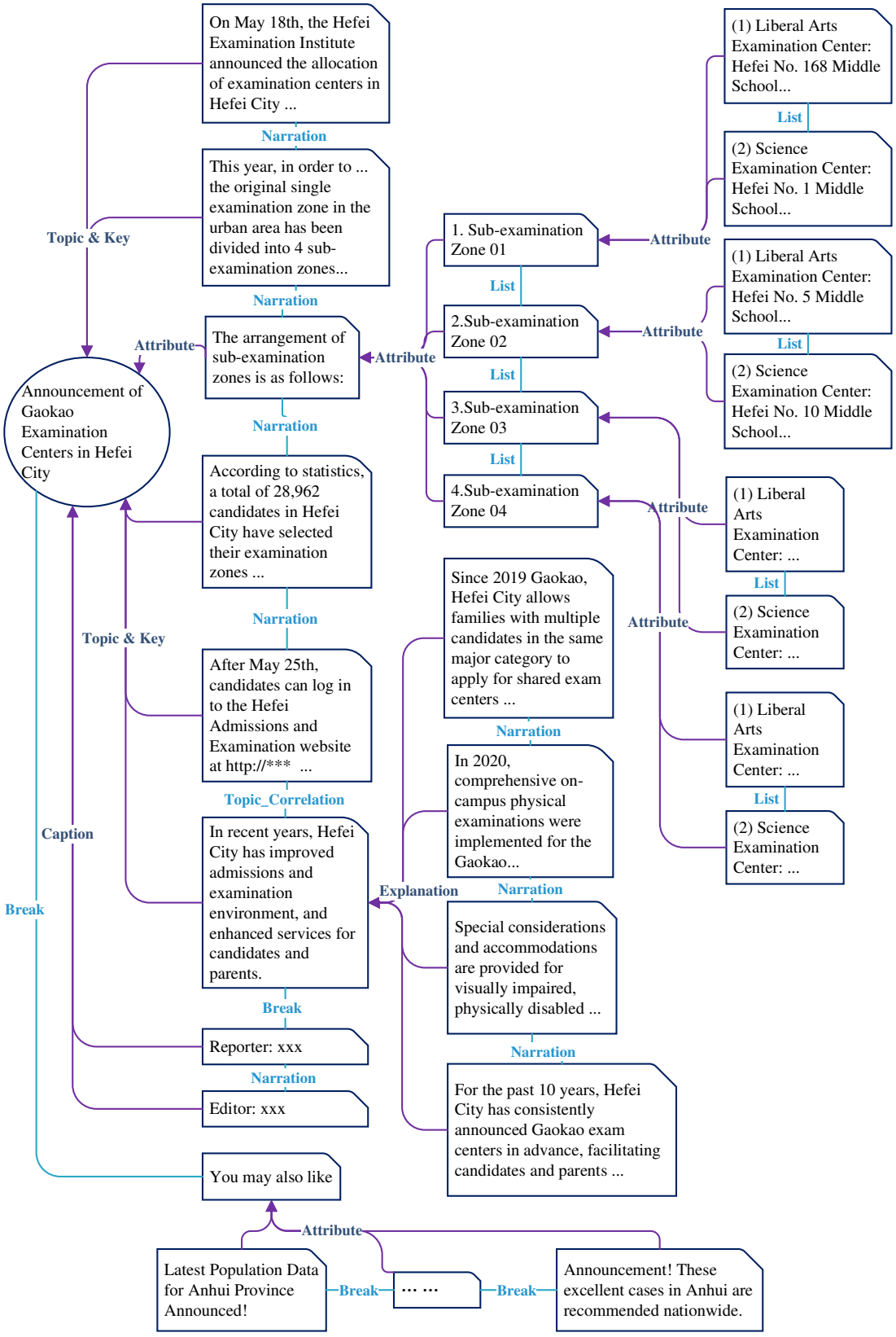

Figure 7: Selected discourse structure example from WEBDOCS dataset. English translation of Figure 6. Some content are omitted with ellipsis for clearer display.

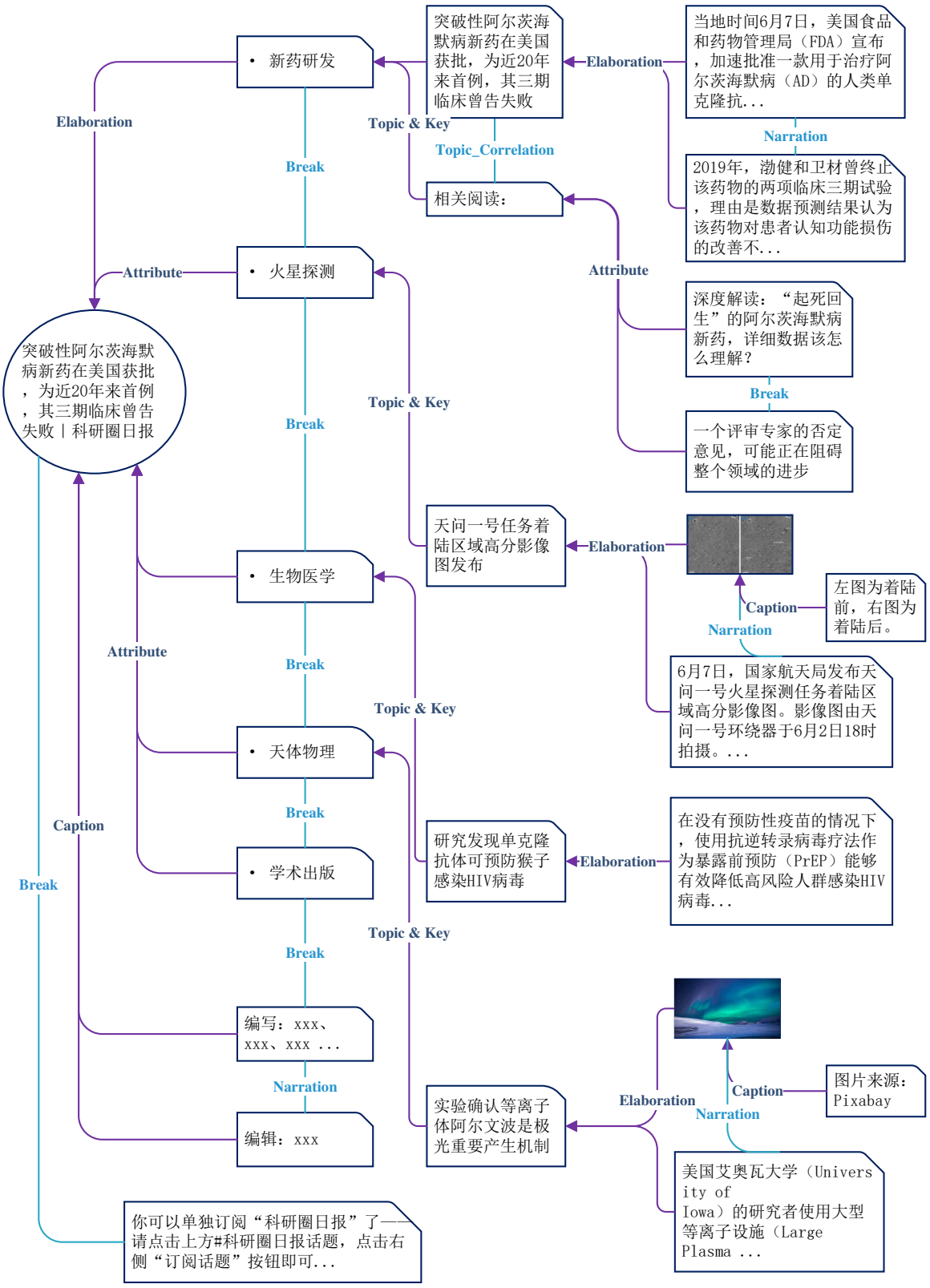

Figure 8: Selected discourse structure example from WEBDOCS dataset. Some content are omitted with ellipsis for clearer display. An English translation version of the same web document discourse structure and web document snippet can be found in Figure 9.

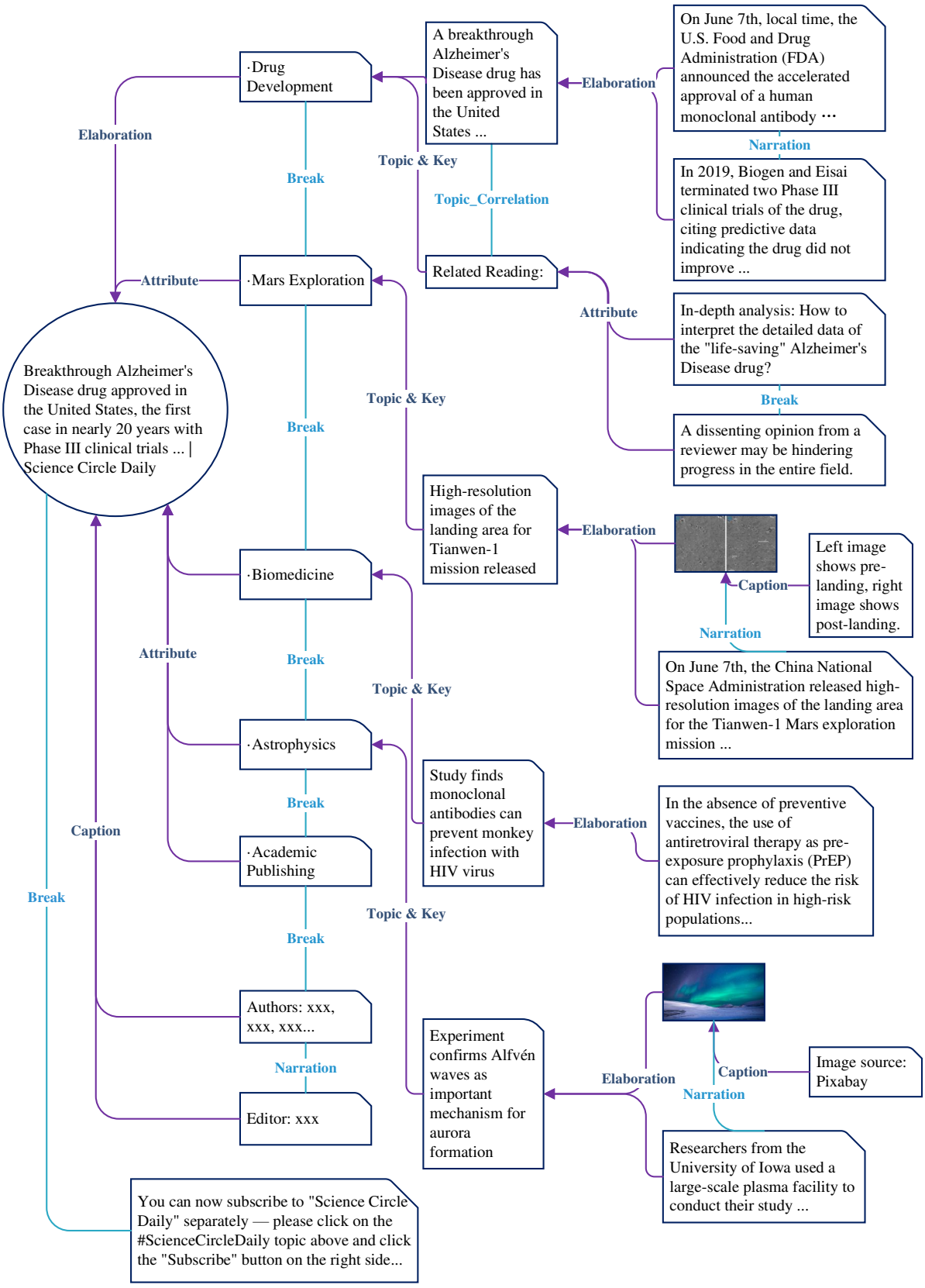

Figure 9: Selected discourse structure example from WEBDOCS dataset. English translation of Figure 8. Some content are omitted with ellipsis for clearer display.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7 (That is the first section after the Conclusion section (6), ACL2023 Template has not given a serial number for it)

☒ A2. Did you discuss any potential risks of your work?
This work is a foundational research which propose a discourse schema for web documents, and is not tied to particular applications. We claim no biased usage for different stakeholders or conflict of interest. For the ethic considerations on our created dataset, such as privacy considerations and malicious words, we show relevant information in the Section "Ethics Statement".

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Introduction is in section 1; Abstract is at the begining of the paper under its title.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4 and Section Appendix B describe our data, that is an annotated data set proposed in our paper, we descirbe the data sourse web site where we collected the data Section 5 descries our re-implemented models for experiments and we cited their authors of original papers who proposed the models We also append our code to the Supplementary Materials

☑ B1. Did you cite the creators of artifacts you used?
Section 4 and Section Appendix B describe our data, that is an annotated data set proposed in our paper, we descirbe the data sourse web site where we collected the data Section 5 descries our re-implemented models for experiments and we cited their authors of original papers who proposed the models

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Section Ethics Statement, we make claim that our collected data is from a publicly available sources and are freely accessible online without copyright constraint to academic use

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In Section Limitation we discuss the harms that may ensue from the limitations of the data collection methodology In Section Ethics Statement we state that the data is sufficiently anonymized

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Ethics Statement (That is the second section after the Conclusion section (6), ACL2023 Template has not given a serial number for it)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix B for Data Collection and Detailed statistics

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 4.1 for details of train/test/dev splits; Appendix B for Data Collection and Detailed statistics

C ☑ Did you run computational experiments?

Section 5 Experiments

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix C Experimental Details

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C Experimental Details

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
In Section 5.2, we report the results which are mean values from 3 runs with different random seed

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
we use PyTorch for computation in deeplearning, and use huggingface for pre-trained language model. we report relevant information In Appendix C Experimental Details

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 4 WebDP: A new benchmark for Web Document Discourse Parsing

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
The full text of instructions given to participants is an annotation guideline document, which is kind of lengthiness and mainly talks about technical details of annotation, so we do not attach it directly to the paper. In any case of need, we can open it by publishing on the internet. The annotation process is conventional and has no risk to participants or annotators, which has been confirmed repeatedly during our preliminary study stage. And all of the annotators we recruited have been well informed about the process before they agree to participate.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section Ethics Statement - 3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section Ethics Statement - 1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section Ethics Statement - 1

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section Ethics Statement - 3