A Topicality-Driven QUD Model for Discourse Processing

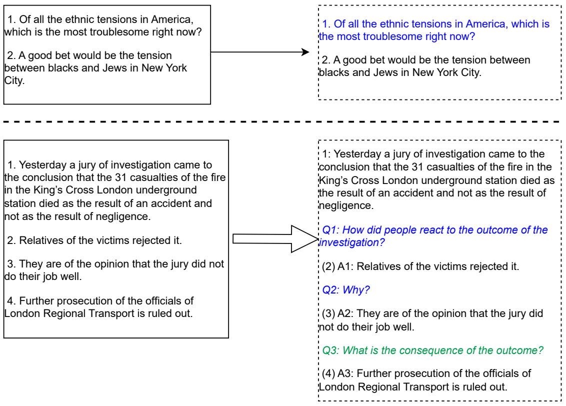
Yingxue Fu[1, 2] *, Mark-Jan Nederhof[1] *, Anaïs Ollagnier[2]
[1]School of Computer Science, University of St Andrews, Scotland, UK
[2]Université Côte d'Azur, Inria, CNRS, I3S, Sophia Antipolis, France
fuyingxue321@gmail.com

Abstract

Question Under Discussion (QUD) is a discourse framework that has attracted growing interest in NLP in recent years. Among existing QUD models, the QUD tree approach (Riester, 2019) focuses on reconstructing QUDs and their hierarchical relationships, using a single tree to represent discourse structure. Prior implementation shows moderate inter-annotator agreement, highlighting the challenging nature of this task. In this paper, we propose a new QUD model for annotating hierarchical discourse structure. Our annotation achieves high inter-annotator agreement: 81.45% for short files and 79.53% for long files of Wall Street Journal articles. We show preliminary results on using GPT-4 for automatic annotation, which suggests that one of the best-performing LLMs still struggles with capturing hierarchical discourse structure. Moreover, we compare the annotations with RST annotations. Lastly, we present an approach for integrating hierarchical and local discourse relation annotations with the proposed model.

1 Introduction

As large language models (LLMs) show impressive performance on various NLP tasks (Wei et al., 2022a; OpenAI, 2023; Wei et al., 2022b), there is an increasing number of studies that try to convert NLP tasks into text generation tasks to leverage the strong generative capability of LLMs (Raffel et al., 2020; Yuan et al., 2021). In computational discourse processing, Question Under Discussion (QUD) (Roberts, 2012; Onea, 2016; von Stutterheim and Klein, 1989) has been attracting growing attention over the years. With this framework, discourse units, typically sentences, are considered as answers to some explicit or implicit questions, which are called QUDs. One QUD may lead to another QUD, and multiple QUDs can jointly contribute to resolving a higher-level QUD[1]. Discourse structure can thus be understood through the relationships between these QUDs. An example is shown in Figure 1. With QUDs reconstructed and added to discourse, the texts remain coherent, and the information conveyed is largely unchanged. Moreover, the relationship between sentences is mirrored by the relationship between QUDs.


1. Of all the ethnic tensions in America, which is the most troublesome right now?

2. A good bet would be the tension between blacks and Jews in New York City.

1. Of all the ethnic tensions in America, which is the most troublesome right now?

2. A good bet would be the tension between blacks and Jews in New York City.

1. Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.

2. Relatives of the victims rejected it.

3. They are of the opinion that the jury did not do their job well.

4. Further prosecution of the officials of London Regional Transport is ruled out.

1: Yesterday a jury of investigation came to the conclusion that the 31 casualties of the fire in the King's Cross London underground station died as the result of an accident and not as the result of negligence.

Q1: How did people react to the outcome of the investigation?

(2) A1: Relatives of the victims rejected it.

Q2: Why?

(3) A2: They are of the opinion that the jury did not do their job well.

Q3: What is the consequence of the outcome?

(4) A3: Further prosecution of the officials of London Regional Transport is ruled out.

Figure 1: The upper example, an excerpt of wsj_2369 from the Penn Discourse Treebank (Webber et al., 2019), shows a text with an explicit question, which is answered by the following sentence. The example below, taken from van Kuppevelt (1995), is more common for written texts. The text with reconstructed QUDs is shown on the right side. As can be seen, *Q2* elaborates on *Q1* regarding people's reactions to the outcome of the investigation, while *Q3* shifts the focus from people's reactions to the consequences of the conclusion.

Similar to the case with canonical discourse frameworks, such as Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Penn Discourse Treebank (PDTB) (Webber et al., 2003),

[*] Equal contribution.

[1] We use "one/a/the QUD", "QUDs" or "questions" to denote specific questions that function as QUDs for discourse units, and "the QUD framework" or "QUD" refers to the general framework.

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 214–230
Aug 25–27, 2025. ©2025 SIGDIAL

different QUD models have been proposed (Fu, 2025), including the QUD-tree approach (Roberts, 2012; Riester, 2019; De Kuthy et al., 2018), the expectation-driven approach (Westera et al., 2020; Kehler and Rohde, 2017) and the dependency-based approach (Ko et al., 2022, 2023; Wu et al., 2024). These models have different focuses: the QUD-tree approach is similar to RST in using a single tree to represent discourse structure; the expectation-driven approach adopts an incremental model, where an analyst (a human reader or a computational system) can only access the preceding context; and the dependency-based approach assumes that each sentence answers a QUD derived from a sentence in the preceding discourse. Among them, the QUD-tree approach and the dependency-based approach take discourse structure into consideration, but as with syntactic parsing, the dependency structure is shallower than the structure obtained with the QUD-tree approach, because the relationships between QUDs are not modeled. For example, in the second example in Figure 1, it cannot capture the parallel relationship between *Q3* and the higher-level QUD formed by *Q1* and *Q2*. Previous research shows that it is challenging to achieve high inter-annotator agreement (IAA) on annotating QUD trees (De Kuthy et al., 2018).

In this work, we aim to advance this line of research by proposing a new QUD model that focuses on hierarchical discourse structure. This model is built on the theoretical proposal by van Kuppevelt (1995), where topicality[2] is the guiding principle for discourse segmentation and discourse structuring. As claimed by van Kuppevelt (1993), this model enables multiple-level discourse analysis advocated by Grosz and Sidner (1986), namely, intentional structure, linguistic structure, and attentional structure, which existing discourse frameworks struggle with (see section 4.1 and section 4.2).

To summarize, our contributions are as follows:

1. We propose a new QUD annotation scheme based on the theoretical framework by van Kuppevelt (1995), and demonstrate high IAA on challenging, naturally occurring texts.

2. We develop an annotation interface for this task, and publicly release the code, sample annotated texts, and annotation guidelines[3].

3. We show preliminary results of automatic annotation using GPT-4, an LLM with strong performance on multiple NLP tasks (OpenAI, 2023).

4. We show an approach for incorporating hierarchical and local discourse relation annotations with the proposed model.

2 Related Work

There exists a substantial body of work on discourse modelling. RST, PDTB, and Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003) are the canonical discourse frameworks. We provide a review of related discourse models to clarify the conceptualization and motivations of our research.

QUD-tree This is the most closely related work to the present study. It can be traced to the model by Roberts (2012). Roberts (2012) argues that people communicate to reach a common ground and QUDs are raised during the process of communication to resolve indeterminacies. When a QUD is accepted, it is put on a stack model of discourse, and discourse participants are committed to answering it. It is popped off the stack when it is answered satisfactorily or when it is determined to be unanswerable. Accordingly, discourse can be modeled by a single tree of QUDs. Riester (2019) develops four constraints for reconstructing QUDs. These constraints also form the theoretical foundation for later studies on evaluating QUD parsing (Wu et al., 2023). Despite the common assumption of using a single tree in modelling discourse, Riester (2019) and De Kuthy et al. (2018) depart from the thesis of Roberts (2012) and allow a weaker connection between QUDs than the criteria defined by Roberts (2012): a QUD is considered valid as long as it is topically related to the QUD at the top of the stack. For discourse segmentation, information-structural units are used as basic units, which can be clauses, and in many cases, they are more fine-grained.

RST RST is the closest related mainstream discourse framework to the QUD-tree approach. Similar to QUD-tree, the first step for RST analysis is typically discourse segmentation. Clauses are the basic discourse units, called elementary discourse

[2] The word "topicality" is used by van Kuppevelt (1995), since discourse structure is assumed to be based on topic-comment relationships, where discourse units are considered as comments while the explicit or implicit questions answered by them constitute topics.

[3] https://github.com/yingxueF/topicality_driven_QUD_sigdial2025

units (EDUs) (Mann and Thompson, 1988). To improve consistency in this step, rules based on syntactic clues are introduced in the creation of the RST Discourse Treebank (RST-DT) (Carlson et al., 2001), which is built on Wall Street Journal (WSJ) articles used in the Penn Treebank (Marcus et al., 1993). EDUs are linked through rhetorical relations and discourse units are grouped in this way recursively until a tree is formed. The relationship between RST and QUD-tree has been investigated in some studies (Shahmohammadi et al., 2023; Ko et al., 2023).

Topic Segmentation Topic segmentation refers to the task of breaking down a text into groupings of smaller segments, where the motivation for the grouping is that the segments deal with a common topic (Jiang et al., 2021; Stede, 2012). Most existing studies on topic segmentation focus on sentence-level topic boundary demarcation (Eisenstein and Barzilay, 2008; Zhang et al., 2023). In comparison, Jiang et al. (2024) take paragraphs as the basic units under the assumption that each paragraph deals with one topic. Consecutive paragraphs may be grouped if they address the same topic, with the document as a whole focusing on a supertopic. Thus, the structure is hierarchical. However, in Jiang et al. (2024)'s study, paragraphs dealing with the same topic are simply grouped, without considering the relationship between paragraphs under the same topic or between topics, and they work with coarse-grained textual units.

3 Theoretical Framework

In the theoretical model proposed by van Kuppevelt (1995), the process of questioning involves three parameters: *feeders*, *topic-constituting questions* and *subtopic-constituting subquestions* (henceforth referred to as subquestions for brevity). Feeders are discourse units that trigger a questioning process, and they serve to introduce indeterminacies, which questions are then raised to resolve. Feeders can be linguistic, such as the opening sentence of a monologue, or non-linguistic, for instance, a door knock that triggers a conversation. The defining characteristic of a feeder is that the discourse unit is either topic-less, not prominent for the current discourse, or occurs in a place of the text where no context is set for it. In the example given in the lower part of Figure 1 (henceforth "London underground example"), the first sentence functions as a feeder because no context is set for it, and accordingly, no QUDs are reconstructed for this discourse unit. Following Roberts (2012), Riester (2019) proposes to add a proxy QUD for feeders — "What is the way things are?". Another example is provided below to illustrate different types of feeders (a modified version of example (10) from van Kuppevelt (1995), henceforth "Mary's holiday example"). There are two feeders, which initiate different discourse topics.

F1: A: Mary is on holiday.

Q1: B: When did she leave?

A1: A: She went to the airport yesterday.

Q2: B: Did she tell you where she was going?

A2: A: She told me she was going to Thailand a month ago.

F2: A: I went to Thailand when I was eight years old.

Q3: B: How did you like the country?

A3: A: My experience with the country was fantastic. If I had more money and time now, I would definitely go there again.

In this dialogue, the first feeder is of the same type as that in the London underground example, while the second feeder belongs to a different type, which originates from a part of the previous topic, i.e.,"going to Thailand".

The introduction of a feeder F initiates a discourse topic, DT, which is typically unfolded as a set of topic-constituting questions $\{TQ_1, TQ_2, TQ_3, \ldots\}$. Hence, a discourse topic overlaps with a topic-constituting question when only one topic-constituting question is induced.

As indicated by van Kuppevelt (1995), topic-constituting questions have an autonomous status in discourse. These higher-order questions impose a restriction on the development of discourse. An unsatisfactory answer to such questions gives rise to a subquestion, which, if not answered satisfactorily, brings up a further subquestion, and so on, until no further subquestions are elicited, leading to a hierarchical structure of questions. Therefore, different from topic-constituting questions, subquestions are contextually induced and do not have an autonomous status in discourse. As shown by the London underground example, Q1 and Q3 are topic-constituting questions, one concerning people's reaction and the other dealing with the consequence of the conclusion. Q2 is induced from A1, which suggests that A1 provides an unsatisfactory answer to Q1, and in this case, Q2 forms a subquestion of Q1.

The process of pursuing a topic-constituting question by raising subquestions is governed by two principles: *the principle of recency* and *the dynamic principle of topic termination*.

The principle of recency suggests that subquestions are evoked by the most recent unsatisfactory answer to a preceding question. In the London underground example, the subquestion Q2 is induced by A1, which is the most recent unsatisfactory answer to Q1. This principle ultimately produces a tree for a topic-constituting question, akin to the stack model proposed by Roberts (2012).

The dynamic principle of topic termination specifies that if an explicit or implicit question is answered satisfactorily, the related questioning process is closed, and the topic carried by the question loses its currency. This explains why in the Mary's holiday example, when A2 is given, the topic associated with Q2 loses its relevance and the next sentence functions as a feeder. A method can be used to determine if the question of a discourse unit is a subquestion of the immediately preceding question: a topic-closing sentence S is added after an answer A_i to a question Q_i. If the immediately following question Q_{i+1} after A_i becomes infelicitous, it means that Q_{i+1} is a subquestion of Q_i.

This method is called *subordination test*. An example of S is "I now understand that...", which forces a preceding topic to be closed. In the London underground example, when a subordination test is applied to test if Q3 is subordinate to Q2:

> F: Yesterday a jury of investigation came to the conclusion that...
>
> Q1: How did people react to the outcome of the investigation?
>
> A1: Relatives of the victims rejected it.
>
> Q2: Why?
>
> A2: They are of the opinion that the jury did not do their job well.
>
> *S*: *I now understand why relatives of the victims rejected it.*
>
> Q3: What is the consequence of the conclusion?
>
> A3: Further prosecution of the officials of London Regional Transport is ruled out.

Q3 is felicitous given S. Therefore, Q3 is not subordinate to Q2.

In the model by Roberts (2012), a QUD that has been answered and popped off the stack is added to the common ground of the discourse model, contributing to a cumulative and evolving common ground that further shapes the development of discourse. In contrast, with the model by van Kuppevelt (1995), questions that have been answered satisfactorily lose their actuality, indicating a transient and dynamic process. Moreover, the model by Roberts (2012) assumes one primary communicative goal—to come to an agreement about the way things are. Thus, discourse is organized around accumulating shared knowledge and reaching consensus. Consequently, discourse is represented by a single hierarchical tree. By comparison, with the model by van Kuppevelt (1995), a discourse topic is defined by a set of topic-constituting questions that form a program about how discourse progresses. It is conceivable that the resulting structure consists of a set of shallower trees.

Apart from discourse structuring, discourse segmentation is also shaped by the organization of questions. With the model, multiple discourse topics can be formed, which is typical for informal conversations, and these topics can be structurally unrelated. At the top level, discourse units of larger sizes address discourse topics. Beneath each discourse topic unit, one or more smaller discourse units can be identified, each associated with topic-constituting questions linked by the common discourse topic. Furthermore, each discourse unit answering topic-constituting questions may be further divided into finer-grained units based on the subquestions they address.

The dynamic principle of topic termination entails topic shifts, which may occur when one topic-constituting question is resolved and discourse progresses to the next topic-constituting question. As mentioned earlier, a feeder initiates a discourse topic. Therefore, two types of topic shifts may be defined: topic shifts under the same feeder (one discourse topic), and topic shifts under successive feeders (more than one discourse topic). Topic shifts under the same feeder happen when a new topic-constituting question arises. The case of topic shifts under successive feeders is more complicated, involving three possibilities: *associated topic shifts*, *non-associated topic shifts* and *topic descending shifts*.

Associated topic shifts refer to the case when the new feeder is contained within a part of the preceding discourse. In the Mary's holiday example, F2 is associated with A2 in the previous discourse. In comparison, non-associated topic shifts occur when the new feeder is irrelevant to the preceding discourse. This is typical in informal conversations, where participants frequently jump from one topic to another. Topic descending shifts form a special case of associated topic shifts, where a preceding subtopic under an old discourse topic becomes the feeder for the new discourse topic, as shown by the following example (originally from van Kuppevelt (1995), edited):

F1 A: Nigel has kicked his dog again.

Q1 B: Why?

A1 A: He had a fight with his wife.

Q2 B: What happened?

F2/A2 A: She had a terrible headache.

Q3 B: How come?

A3 A: She had been addicted to drugs for a while.

Q4: B: Didn't she say she'll quit this?

A4: A: Yeah, but you know her.

Hence, the model allows for capturing incoherence caused by topic shifts, which are common in dialogues.

4 Positioning Against Existing Frameworks

4.1 Multi-level Analysis

van Kuppevelt (1993) argues that the model enables multi-level analysis, which forms a challenge for RST analysis, since only one relation label, either informational or intentional, is allowed between two connected discourse units, resulting in insufficient representation of information at both levels (Moore and Pollack, 1992). We take the multi-level discourse model by Grosz and Sidner (1986) as an exemplary framework for discussion. In linguistic structure, it has been discussed in the above section how topicality functions as the guiding principle of discourse segmentation. The intentional structure is represented by the restrictions formed by the discourse topic and the topic-constituting questions subsumed within it. These questions define what is to be communicated adequately for a discourse to come to an end. Each topic-constituting question may entail different ways of subquestioning to effect changes in a discourse participant's mind (van Kuppevelt, 1993). For the attentional structure, as discussed above, subquestions are contextually induced, and as long as new subquestions are raised, the associated higher-level questions persist in focus. When subquestions are answered satisfactorily, they lose their actuality.

4.2 Insights from RST Parsing

RST posits that the same set of relations can be applied to textual spans of arbitrary sizes (Stede, 2008; Taboada and Mann, 2006). However, studies on RST parsing reveal that discourse relations are distributed differently at intra-sentential and inter-sentential levels (Joty et al., 2013). Williams and Power (2008) present evidence that some rhetorical relations tend to occur higher up the RST tree than others, spanning over large segments of text. This kind of relations is more challenging for automatic systems, as shown by Liu et al. (2023). Since a single tree is used to represent the structure of an entire text, an RST tree can become quite deep, especially for longer texts. As Huber et al. (2022) point out, tree aggregation at higher levels differs significantly from that at lower levels. While EDU-level tree aggregation primarily relies on local syntactic and semantic features, higher-level tree aggregation utilizes global features, such as topic shifts intended by the author to achieve specific communicative goals. Therefore, recent studies on RST parsing, including those by Huber et al. (2022), Jiang et al. (2021), and Peng (2023), focus on the more challenging task of macro-level RST parsing, i.e., RST parsing above the sentence level. In the modified RST model by Knott et al. (2000), coherence at higher levels is achieved through a different mechanism than at lower levels.

The above insights align with the model here, which postulates different types of coherence at higher and lower levels. Higher-level topic-constituting questions are connected by the same discourse topic, while subquestions are hierarchically organized under topic-constituting questions.

4.3 Insights from PDTB *NoRel*

The PDTB *NoRel* relation receives little attention in existing studies. As RST analysis provides full coverage of the text, everything in the text has to be a part of the analysis. Therefore, when no relations are considered to hold between two arguments in PDTB, the RST annotation may contain a relation[4]. Table 1 shows some examples on how instances of PDTB *NoRel* relation are annotated in RST. The corresponding RST annotations are linearized using the method proposed by Braud et al. (2016) for compact representation.

It can be observed that the RST spans that corre-

[4] in personal communication with Bonnie Webber (Jan, 2024)

PDTB	RST	Source
“arg1”: “The company went public earlier this month, offering 1,745,000 shares of common stock at $15 a share.”, “arg2”: “Giant has interests in cement making and newsprint.”	NS-Elaboration-Additional(NS-Elaboration-Additional(7, NN-Temporal (8, 9)), 10)	wsj_0695
“arg1”: “The appointment increased the number of directors to 10, three of whom are company employees.”, “arg2”: “Simpson is an auto parts maker.”	NS-Elaboration-Additional(NS-Elaboration-Additional(1, NS-Elaboration-Set-Member(2, 3)), 4)	wsj_1119

Table 1: Examples of PDTB *NoRel* relation and their corresponding RST annotations. Numbers denote EDUs in RST annotation, where red denotes EDUs under the span of “arg1” and blue indicates EDUs in the span of “arg2”. In the linearized format, Nuclearity-SenseLabel is displayed before the span, and hierarchical relationship is indicated by nested brackets.

spond to the two arguments linked by PDTB *NoRel* are located in different subtrees, and they are combined with other elements to form a subtree before being connected with the span corresponding to the other argument in PDTB-style annotation. For instance, in the example of wsj_0695, the RST span (8, 9), corresponding to argument 1, is connected with EDU 7 first, and the whole subtree is connected with EDU 10 through an *Elaboration-Additional* relation. In the original text, the span formed by (8, 9) is a sentence, and EDU 10 is an adjacent sentence. Based on the annotation procedure of PDTB, annotators will need to give a label to represent the relation between the two consecutive sentences. The instances shown here are consistently labeled as *NoRel* in PDTB, which does not contradict the information shown in RST annotation (see Appendix A for more examples). In this sense, this label may be considered as a by-product of the local-focused approach adopted by PDTB. High-level analysis seems a beneficial complement to account for some annotations. On the other hand, *NoRel* corresponds to and could be helpful in identifying topic transitions[5].

5 Annotation

5.1 Text Selection and Preprocessing

The main purpose of the annotation is to see how reliably the model can be applied to naturally occurring data. To facilitate comparison, the intersection of WSJ articles from RST-DT and PDTB 3.0 (Webber et al., 2019) is utilized. However, some articles only report increases/decreases in stock prices or transaction volumes, such as wsj_0649 (Appendix B). The QUDs would be along the lines of “What about next?”[6]. Therefore, only articles describing events or with a clear storyline are chosen.

Five short files (≤ 10 sentences) and three long files (> 10 sentences) are selected (Appendix C). Given that WSJ articles are non-trivial for average readers, and existing studies have shown that QUD annotation tasks are challenging (De Kuthy et al., 2018; Westera et al., 2020), a small number of files may be deemed acceptable for our purpose, similar to the study by De Kuthy et al. (2018), where three documents are annotated to test what IAA is achievable.

Since the texts are originally divided into paragraphs, they are processed using the treebank module of the NLTK library (Bird et al., 2009) to retrieve the gold sentence-level splits. The sentences are then numbered sequentially for easy reference by annotators. Manual examination reveals that topic shifts do not necessarily overlap with paragraph boundaries.

5.2 Manual Annotation

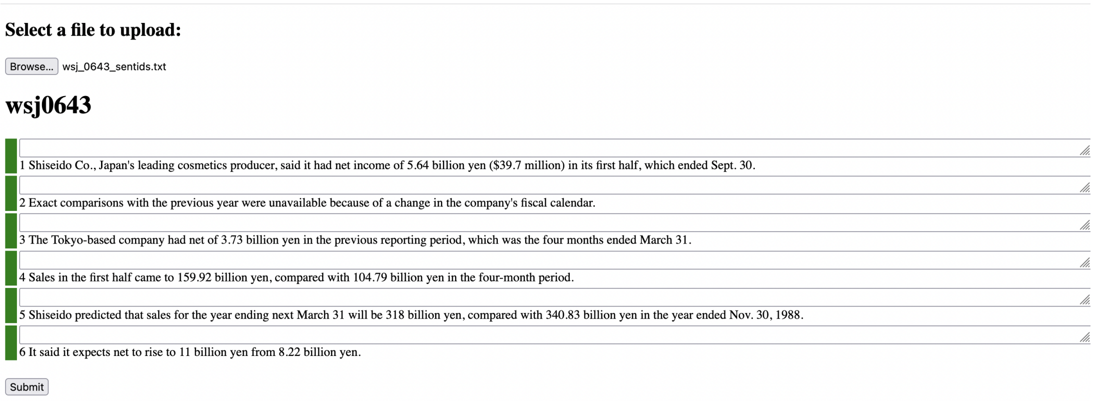
In addition to an author (Annotator A), the annotators (Annotators B and C) are two experts in computational linguistics, but it is the first time for them to annotate a high-level linguistic phenomenon[7]. Annotator B only managed to participate in annotating short files due to other commitments. Therefore, short files are annotated by Annotators A and B, and the more challenging long files are annotated by Annotators A and C. The annotation project has been reviewed and approved by the school’s ethics committee. We develop an annotation interface, and a screenshot is shown in Figure 2.

The trial stage consists of two steps. The first step involves understanding the annotation guide-

[5] As an anonymous reviewer pointed out, *NoRel* annotations exist only within paragraphs in PDTB. As most topic transitions in a newspaper are likely to happen between paragraphs, only a very small fraction of cases will be captured by *NoRel* annotations. We argue that topic transitions within paragraphs are more challenging to detect because of lack of clear signals.

[6] in communication with Amir Zeldes (November, 2024)

[7] Initially, undergraduates without experience in computational linguistics volunteered but withdrew after the trial stage.



Select a file to upload:

wsj_0643_sentids.txt

wsj0643

1 Shiseido Co., Japan's leading cosmetics producer, said it had net income of 5.64 billion yen ($39.7 million) in its first half, which ended Sept. 30.

2 Exact comparisons with the previous year were unavailable because of a change in the company's fiscal calendar.

3 The Tokyo-based company had net of 3.73 billion yen in the previous reporting period, which was the four months ended March 31.

4 Sales in the first half came to 159.92 billion yen, compared with 104.79 billion yen in the four-month period.

5 Shiseido predicted that sales for the year ending next March 31 will be 318 billion yen, compared with 340.83 billion yen in the year ended Nov. 30, 1988.

6 It said it expects net to rise to 11 billion yen from 8.22 billion yen.

Figure 2: A screenshot of the annotation interface. Each sentence has a blank text box for the annotation of its QUD. The code for the interface, a demo, and annotated samples are released on github.

lines, which takes about an hour. In the second step, all the annotators work on two trial files, which takes around 1.5 hours. Then the annotators compare the annotated files to see if there are any misunderstandings of the task and how their agreement can be improved. After this stage, the annotators work on the annotation task independently.

The method proposed by De Kuthy et al. (2018) is adopted for computing IAA scores in structure annotation, which is based on the approach devised by Marcu et al. (1999) for measuring IAA in RST tree construction. Since the method has been explained by De Kuthy et al. (2018), we only show an example to elucidate the details in Appendix D. The Cohen's Kappa (κ) values (Cohen, 1960) are calculated. It is on average 0.8145 for short files and 0.7953 for long files.

Table 2 shows IAA score for each of the files[8].

Types	Filenames	Number of Cells	κ
Short files	wsj_1381	66	0.9533
	wsj_1985	55	0.7239
	wsj_2313	66	0.7664
Long files	wsj_0601	435	0.6777
	wsj_1184	325	0.8945
	wsj_2339	231	0.8136

Table 2: IAA scores for manual annotation.

Unsurprisingly, higher IAA scores can be achieved on short texts, but the results on long texts are close. While the IAA scores are much higher than those shown by De Kuthy et al. (2018) (0.52 on average for discourse structure), the results are not directly comparable, because the two annotation efforts differ in complexity, genre and lengths of selected texts, granularity in discourse segmentation, and structural constraints.

As free texts are used, it is a challenge to measure IAA on identified questions. De Kuthy et al. (2018) argue that it is futile to compare QUDs given by different annotators using string matching, since languages allow innumerable ways of expressing a question. If the identified QUDs entail the same discourse structure and information structure, the evaluation of the surface form of QUDs can be discounted. Nevertheless, following Westera et al. (2020), we show some quantitative results.

Cosine similarity scores between sentence embeddings of annotated questions are computed. The Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) model is utilized for obtaining sentence embeddings. The average cosine similarity score is 0.4869. For a subjective task that involves high variability in question formulation, this value can be considered moderate[9]. The Jaccard similarity metric is also used, which is based on the percentage of intersection over union of two tokenized sentences. The average similarity score is 0.1970, indicating a low lexical overlap of questions between annotators. Manual examination reveals that in some cases, the Jaccard similarity score is low, but the cosine similarity metric captures semantic similarity correctly. Table 3 shows examples of question similarity measured by the two metrics.

To determine if span size influences IAA on in-

[8] Among the eight files, two short files were used in the pilot annotation stage and the remaining six files are included in computing IAA scores.

[9] We don't have a baseline to evaluate what this cosine similarity score actually means, especially considering that question generation is tied with the results of discourse structuring. A weak model, such as GPT-2, can be used to generate questions and the results can be used for comparison. While this is useful for comparing different LLMs, the task here is IAA achievable by human annotators.

Annotator1	Annotator2	Jaccard	Cosine
"Who are those involved in the partnership?"	"Who are involved in the partnership?"	0.8571	0.9733
"How does GE interpret its activities?"	"How does GE describe its activities?"	0.7143	0.9418
"What influence does Imperial expect the suit to have?"	"What does Imperial expect following the suit?"	0.6000	0.9186
"Why is it the case that the federal judiciary one of the last bastions of the generalist?"	"Why have there been generalists in the federal judiciary?"	0.2222	0.8746
"What is Mr. Orr's opinion?"	"Has there been a response from Mr. Orr?"	0.2000	0.8620
"What does this suggest?"	"Is it common that the government takes such actions?"	0.0000	0.0462
"What are political reasons?"	"Was this action expected and why?"	0.0000	0.1051
"How does the government interpret GE's activities?"	"What is this wrongdoing and in which context did it occur?"	0.0000	0.1418
"Why is the deal politicized?"	"What are the prospects for the acquisition going ahead?"	0.0833	0.1483
"What action has been taken by industry?"	"Who has expressed opinions about the judges that should be hired?"	0.0588	0.1646

Table 3: Examples of questions given by different annotators over matched spans. Ten question pairs are shown: five with the highest and five with the lowest cosine similarity scores.

ferred questions, the Pearson correlation between span sizes and cosine similarity scores is computed: -0.105 (p-value = 0.243), and with Jaccard similarity scores, it is 0.117 (p-value = 0.195). Both results indicate no significant relationship. This suggests that the results are not influenced by the distance between spans, which may demonstrate that the task is not dependent on local linguistic cues.

5.3 Automatic Annotation

One of the applications of LLMs is data annotation (Ding et al., 2023). This section shows a set of experiments on automatic annotation with LLMs.

The model *gpt-4-turbo* (OpenAI, 2023) is used. As the task is complex, involving reasoning for high-level linguistic phenomena, the Chain-of-Thought (CoT) prompting method (Wei et al., 2022b) is adopted. Trial experiments show that few-shot prompting generates better performance. Therefore, a three-shot CoT prompting approach is employed (see Appendix E for the prompt template).

The input to the model is a list of sentences, as in the case of human annotators, and the output is the parent of each sentence and the generated questions. As both the input and output are long, the max token parameter is set to 3000. The temperature is set to 0 to increase determinism in the output, and only one reply is needed.

Since the performance of the model is below expectations for long files (too shallow structure), only Cohen's κ showing IAA between GPT-4 and Annotator A on short files is computed. It is 0.5337, indicating that human annotators can achieve much higher agreement with each other than with GPT-4. The detailed IAA scores per file are shown in Appendix F. Manual examination of GPT-4's annotation for wsj_2313, the file with the lowest IAA score, reveals that there is still room for improvement in discourse understanding for GPT-4 (see Appendix G for details).

6 Comparison with RST Trees

Owing to the similarity in encoding hierarchical discourse structure, we make a comparison between the annotated QUD trees and RST trees. As sentences are used as the basic unit in the present study, only RST trees above the sentence level are extracted. When assuming a discourse topic at the top, the structure produced by the QUD model may be closer to RST. Therefore, when comparing the heights of RST and QUD trees, the heights of QUD trees are computed under the assumption that an overall discourse topic is present.

Table 4 presents the results of comparison between QUD and RST trees. RST trees are generally taller than QUD trees. Discourse participants tend to address a QUD as soon as it is recognized, which helps reduce the cognitive load. This may explain why the QUD trees are typically shallower

Filenames	RST(Tree-h)	QUD(Tree-h)	Matching
wsj_1381	5	3	0.3750
wsj_1985	7	3	0.4706
wsj_2313	4	4	0.5000
wsj_0601	12	8	0.4583
wsj_1184	8	6	0.4500
wsj_2339	7	5	0.5625

Table 4: Comparison between RST and QUD trees. The last column shows the ratio of spans where both an RST relation and a QUD exist, without considering leaf nodes.

than trees partly based on semantic links. The last column shows that about half of the non-terminal nodes in the long files (below) are annotated with both RST relations and QUDs, while the ratio is lower for short files (above). Manual examination reveals structural differences between RST and QUD structures in short files.

Even though there are spans that are annotated with an RST relation and a QUD in parallel, it is difficult to draw any conclusions on the relationship between RST relations and QUDs, because questions can be formulated in different ways. For example, as shown in Table 5, *why*-questions can be associated with *Background*, *Explanation* and *Elaboration* relations. In contrast, Westera et al. (2020) show a statistically significant correlation between *why*-questions and causal relations. However, the questions inferred using their approach are closer to PDTB-style local relations. This may suggest that, even within the general QUD framework, questions identified through different approaches are not inherently similar.

Questions	RST Relations
"Why is the case an example of poor legal reasoning by judges who lack patent litigation experience?"	Elaboration
"Why is the government raising this obstacle?"	Background
"Why is the government's action unusual?"	Explanation

Table 5: Questions and RST relations over the same spans.

7 Integrating Local and Hierarchical Annotations

Most existing studies on QUD treat sentences as the basic unit (Fu, 2025). An exception is the research by Riester (2019), which allows for the inclusion of sub-sentential units. However, a significant proportion of these units do not possess their own QUDs, leading to cases of QUDs with joint answers (Appendix H). Since sentences are the basic discourse units, the current QUD model does not consider intra-sentential relations. One reason is that it is often challenging to identify topics for sub-sentential units. Following the line of research on integrating different perspectives of discourse annotation (Riester et al., 2021; Fu, 2022; Braud et al., 2024; Liu et al., 2024; Fu, 2024; Zeldes et al., 2025), we show a method of integrating local and hierarchical discourse annotations to enable a full question answering (QA) approach for discourse processing, with the proposed model for processing inter-sentential relations and the method proposed by Pyatkin et al. (2020) for handling intra-sentential relations.

The method by Pyatkin et al. (2020) has been shown to be applicable to naturally occurring texts. However, their experiments involve a preprocessing step of target extraction, which might introduce biases for our task, as discourse relations are not necessarily inferred based on the relationships between entities. Therefore, we adopt the criteria of determining intra-sentential arguments in PDTB 3.0 (Webber et al., 2019) as a first step to identify discourse units at the intra-sentential level. Similar to Pyatkin et al. (2020), we allow more than one question to be raised for a pair of intra-sentential arguments. Therefore, given a list of sentences, intra-sentential level processing involves two subtasks: identifying the argument pairs within each sentence (more than one pair may be possible), and converting these argument pairs into QA pairs using the question templates provided by Pyatkin et al. (2020).

8 Conclusions

We introduce and implement the topicality-driven QUD model proposed by van Kuppevelt (1995) for annotating hierarchical discourse structure. We position the model in existing discourse frameworks and analyze its properties. With this model, high IAA can be achieved on challenging, naturally occurring texts. We compare the annotations with RST annotations to have a better understanding of the generated hierarchical structures and the correlation between QUDs and RST relations. Since this model takes sentences as the basic discourse unit, the method proposed by Pyatkin et al. (2020) is adopted for intra-sentential level processing, thereby enabling a full QA approach for discourse processing.

Limitations

Crowdsourcing platforms such as Amazon Mechanical Turk may be used to create a larger corpus. However, this plan is only practical if the annotation scheme is demonstrated to be simple enough for lay annotators, and sufficient funding for the annotation project is secured. With just a small number of documents being analyzed, this paper is more a proof of concept for a new annotation scheme, and does not present an actually usable corpus.

Another limitation is the lack of a baseline method in question similarity evaluation. For annotations using free-form texts, this is a challenging task, which accounts for the approach adopted in prior studies (De Kuthy et al., 2018).

Acknowledgments

We thank Fahrurrozi Rahman for helping with data annotation and an anonymous reviewer for insightful comments. This work was supported in part by a Santander Development Fund.

References

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

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Chloé Braud, Barbara Plank, and Anders Søgaard. 2016. Multi-view and multi-task training of RST discourse parsers. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1903–1913, Osaka, Japan. The COLING 2016 Organizing Committee.

Chloé Braud, Amir Zeldes, Laura Rivière, Yang Janet Liu, Philippe Muller, Damien Sileo, and Tatsuya Aoyama. 2024. DISRPT: A multilingual, multi-domain, cross-framework benchmark for discourse processing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4990–5005, Torino, Italia. ELRA and ICCL.

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*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. QUD-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Yingxue Fu. 2024. Automatic alignment of discourse relations of different discourse annotation frameworks. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 27–38, Torino, Italia. ELRA and ICCL.

Yingxue Fu. 2025. A survey of QUD models for discourse processing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1722–1732, Albuquerque, New Mexico. Association for Computational Linguistics.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Patrick Huber, Linzi Xing, and Giuseppe Carenini. 2022. Predicting above-sentence discourse structure using distant supervision from topic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10794–10802.

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Fang Kong. 2021. Hierarchical macro discourse parsing based on topic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13152–13160.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2024. Advancing topic segmentation and outline generation in Chinese texts: The paragraph-level topic representation, corpus, and benchmark. In *Proceedings of the*

2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 495–506, Torino, Italia. ELRA and ICCL.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 486–496, Sofia, Bulgaria. Association for Computational Linguistics.

Andrew Kehler and Hannah Rohde. 2017. Evaluating an expectation-driven question-under-discussion model of discourse interpretation. *Discourse Processes*, 54(3):219–238.

Alistair Knott, Jon Oberlander, Michael O'Donnell, and Chris Mellish. 2000. Beyond elaboration: The interaction of relations and focus in coherent text. In *Text Representation: Linguistic and Psycholinguistic Aspects, chapter 7*, pages 181–196. John Benjamins.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022. Discourse comprehension: A question answering framework to represent sentence connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11752–11764, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, Wesley Scivetti, Yilun Zhu, Shabnam Behzad, Lauren Elizabeth Levine, Jessica Lin, Devika Tiwari, and Amir Zeldes. 2024. GDTB: Genre diverse data for English shallow discourse parsing across modalities, text types, and domains. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12287–12303, Miami, Florida, USA. Association for Computational Linguistics.

Yang Janet Liu, Tatsuya Aoyama, and Amir Zeldes. 2023. What's hard in English RST parsing? Predictive models for error analysis. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 31–42, Prague, Czechia. Association for Computational Linguistics.

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

Daniel Marcu, Estibaliz Amorrortu, and Magdalena Romera. 1999. Experiments in constructing a corpus of discourse trees. In *Towards Standards and Tools for Discourse Tagging*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Johanna D. Moore and Martha E. Pollack. 1992. A problem for RST: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4):537–544.

Edgar Onea. 2016. Potential questions at the semantics-pragmatics interface. In *Potential Questions at the Semantics-Pragmatics Interface*. Brill.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Siyao Peng. 2023. *Cross-Paragraph Discourse Structure in Rhetorical Structure Theory Parsing and Treebanking for Chinese and English*. Georgetown University.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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.

Arndt Riester. 2019. Constructing QUD trees. In *Questions in discourse*, pages 164–193. Brill.

Arndt Riester, Amalia Canes Nápoles, and Jet Hoek. 2021. Combined discourse representations: Coherence relations and questions under discussion. In *Proceedings of the First Workshop on Integrating Perspectives on Discourse Annotation*, pages 26–30, Tübingen, Germany. Association for Computational Linguistics.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Sara Shahmohammadi, Hannah Seemann, Manfred Stede, and Tatjana Scheffler. 2023. Encoding discourse structure: Comparison of RST and QUD. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 89–98, Toronto, Canada. Association for Computational Linguistics.

Manfred Stede. 2008. Disentangling nuclearity. *'Subordination' versus 'Coordination' in Sentence and Text: A cross-linguistic perspective*, 98:33.

Manfred Stede. 2012. *Discourse processing*, volume 15. Morgan & Claypool Publishers.

Maite Taboada and William C Mann. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies*, 8(3):423–459.

Jan van Kuppevelt. 1993. Intentionality in a topical approach of discourse structure. In *Intentionality and Structure in Discourse Relations*.

Jan van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(1):109–147.

Christiane von Stutterheim and Wolfgang Klein. 1989. Referential movement in descriptive and narrative discourse. In *North-Holland Linguistic Series: Linguistic Variations*, volume 54, pages 39–76. Elsevier.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. *Philadelphia, University of Pennsylvania*.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1118–1127, Marseille, France. European Language Resources Association.

Sandra Williams and Richard Power. 2008. Deriving rhetorical complexity data from the RST-DT corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? Salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987, Miami, Florida, USA. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72.

Qinglin Zhang, Chong Deng, Jiaqing Liu, Hai Yu, Qian Chen, Wen Wang, Zhijie Yan, Jinglin Liu, Yi Ren, and Zhou Zhao. 2023. Mug: A general meeting understanding and generation benchmark. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A More Instances of *NoRel*

Table 6 shows examples of PDTB *NoRel* instances and their corresponding RST annotations.

1	wsj_0606	a.	“arg1”: “Contract details, however, haven’t been made public.”, “arg2”: “The complex is to be located in Batangas, about 70 miles south of Manila.”	NS((4-9), (NS((10-11), 12)), NS(NS(13, (14-16)), (17-20))
2	wsj_0633	a.	“arg1”: “The stars do that themselves.”, “arg2”: “NBC News has produced three episodes of an occasional series produced by Sid Feders called ‘Yesterday, Today and Tomorrow,’ starring Maria Shriver, Chuck Scarborough and Mary Alice Williams, that also gives work to actors.”	NS((61-65), NS((66-67), 68)), NS(NS(NS(NS(69, 70))), 71), 72)
		b.	“arg1”: “Entertainment shows tend to cost twice that.”, “arg2”: “Re-enactments have been used successfully for several seasons on such syndicated ‘tabloid TV’ shows as ‘A Current Affair’, which is produced by the Fox Broadcasting Co. unit of Rupert Murdoch’s News Corp.”	NN((79-84), 85), NS((86-95), (96-141))
		c.	“arg1”: “ ‘I don’t talk about my work,’ he says.”, “arg2”: “The president of CBS News, David W. Burke, didn’t return numerous telephone calls.”	NN-List(NS(189, (190-191)), 192)
3	wsj_0635	a.	“arg1”: “Ciba Corning, which had been a 50-50 venture between Basel-based Ciba-Geigy and Corning, has annual sales of about $300 million, the announcement said.”, “arg2”: “Terms of the transaction weren’t disclosed.”	NS-Elaboration-Additional(4-7, 8)
		b.	“arg1”: “Terms of the transaction weren’t disclosed.”, “arg2”: “Ciba Corning makes clinical diagnostics systems and related products for the medical-care industry.”	NS(NS((1-3), (4-8)), 9)
4	wsj_0636	a.	“arg1”: “Last week CBS Inc. canceled ‘The People Next Door.’ ”, “arg2”: “NBC’s comedy had aired Wednesdays at 9:30 p.m. and in five outings had drawn an average of only 13.2% of homes, lagging behind the Jamie Lee Curtis comedy ‘Anything But Love’ on ABC and CBS’s one-hour drama ‘Jake and the Fatman’.”	NS((1-2), NS(3, 4)), NN(5, (6-7))
5	wsj_0671	a.	“arg1”: “That’s a taxable-equivalent yield nearly three percentage points more than the current yield on 30-year Treasury bonds.”, “arg2”: “How quickly things change.”	NS-Comment(NS(20, NS(21, 22)), NS(23, (24-31)))
		b.	“arg1”: “A spokesman for Campeau called the rumors ‘ridiculous’.”, “arg2”: “Most investment-grade bonds fell 3/8 to 1/2 point.”	NS-Elaboration-General-Specific(NS((167-168), 169), 170)
6	wsj_0692	a.	“arg1”: “If we want meaningful priorities, we must understand the trade-offs they imply before we make commitments.”, “arg2”: “Strategy is not a separate event in an idealized sequence of discrete events;”	NN-Contrast(NN(174, NS((175-176), (177-180))), NN(181, (182-184)))
		b.	“arg1”: “Mr. Spinney is a permanent Pentagon official.”, “arg2”: “This is a condensed version of an essay that will appear in the January issue of the Naval Institute Proceedings.”	NN-List(185, NS(186, 187))
7	wsj_1124	a	“arg1”: “Revenue more than doubled to $2.62 billion from $1.29 billion.”, “arg2”: “A Salomon spokesman said its stock, bond and foreign exchange trading, as well as its investment banking operations, were mostly responsible for the earnings jump.”	NS(NS(NS(1, 2), SN(3, NN(4, 5))), SN(6, 7))

Table 6: Some (non-exhaustive) examples of RST annotations where a PDTB *NoRel* relation is present. Corresponding RST-style annotations are shown in the linearized format proposed by Braud et al. (2016). Some arguments in PDTB-style annotation are formed by more than one EDU, such as the third example in wsj_0633 and wsj_0635. The details inside the spans are not the focus of the study, hence not expanded. It can be observed that most of the time, the corresponding RST spans are in different subtrees, but PDTB-style annotation cannot capture this type of links owing to its focus on local relations. When the two spans corresponding to two arguments are not in the same subtree, the annotations are not shown in details, since the subtree that one span is in may involve large spans of texts. Structural differences are highlighted.

B Text of wsj_0649

The original text of wsj_0649 is shown below:

> Some parts are masked for copyright protection.
>
> 1 Sharp Corp., Tokyo, said net income in its first half rose ***% to *** billion yen ($*** million) from *** billion yen a year earlier.
>
> 2 The consumer electronics, home appliances and information-processing concern said revenue in the six months ended Sept. 30 rose ***% to *** billion yen from *** billion yen.
>
> 3 Sales of information-processing products and electric parts increased a strong ***% to *** billion yen from *** billion yen and accounted for ***% of total sales.
>
> 4 In audio equipment, sales rose **% to ** billion yen from ** billion yen.
>
> 5 Sales of ** appliances were flat, and sales of *** equipment declined **.
>
> 6 Sharp projected sales for the current year ending ** at ** trillion yen, a *% increase the previous fiscal year.
>
> 7 It said it expects net to rise **% to ** billion yen.

C Details of Selected Data for Annotation

Table 7 shows file names of WSJ articles selected for annotation.

Types	Filenames	Number of Sents
Trial files	wsj_0621	8
	wsj_0653	8
Short files	wsj_1381	11
	wsj_1985	10
	wsj_2313	11
Long files	wsj_0601	29
	wsj_1184	25
	wsj_2339	21

Table 7: Files used in the annotation project. Two short files are used in the trial stage.

D IAA Computation

This section shows how the annotated structure is converted to a matrix for computing Cohen's κ.

For an annotated structure like the following:

```
1-4: ******?
    1: ******?
    2-4: ******?
        2-3: ******?
            2:
            3:
        4: ******?
5-7: ******?
    5: ******?
        6: ******?
            7: ******?
8: ******?
```

the results are shown in Table 8. The method involves checking if a QUD exists for a possible span, and converting a hierarchical structure into a matrix filled with binary values indicating the presence/absence of a QUD. For QUDs with joint answers, such as 2 and 3, only the whole span 2-3 is considered as having a QUD. For subordinating QUDs, such as 5 and 6, it is considered that QUDs are present for the span 5-6 and for 6, since the span involves both 5 and the QUD for 6 as its child nodes, and there is a QUD for 6 involved in the span. For parallel QUDs, it is considered that each node has a QUD, and there is a higher-level QUD for the nodes, similar to the case of "d-trees" by De Kuthy et al. (2018). The cell (8, 8) is 1 because 8 is a node parallel to spans 1-4 and 5-7, under the assumption of an overall discourse topic. If the value at (8, 8) is 0, a piece of information may be lost when higher-level topic-constituting questions are extracted to form the discourse topic. As the cell (1, 1) always has the question "What is the way things are?", it is filtered out in the computation of IAA.

	1	2	3	4	5	6	7	8
1	1	0	0	1	0	0	0	0
2		0	1	1	0	0	0	0
3			0	0	0	0	0	0
4				1	0	0	0	0
5					0	1	1	0
6						1	1	0
7							1	0
8								1

Table 8: The matrix obtained with the method proposed by De Kuthy et al. (2018).

When the annotated structures are converted into this format, the Cohen's κ can be calculated to obtain a statistical measure of IAA for categorical data. As is mentioned by De Kuthy et al. (2018), for a text with n units, the total number of cells is $n * (n + 1)/2$, with the lower half of the matrix discarded.

E Prompt Template for QUD Annotation

Table 9 shows the prompt template for automatic annotation using GPT-4.

You are a language expert at analyzing higher-level text structure. You will analyze some texts based on a discourse model, for which topicality is the general organizing principle of discourse structure, which is a hierarchy of the overall discourse topic, topics under the discourse topic, and subtopics defined for discourse units, which are sentences. The subtopic associated with a discourse unit is provided by the explicit or implicit question the discourse unit answers relative to the overall discourse topic and the dominating topic. Questions can be divided into subquestions, with a purpose of getting more information or resolution of a discrepancy. A (sub)topic constituted by a (sub)question is continued as long as subquestions of that question arise in discourse. If an explicit or implicit (sub)question is answered satisfactorily, the questioning process associated with it comes to an end. Thus, a hierarchy is built recursively until each discourse unit is analyzed. Questions that constitute a topic do not arise without a cause, and a discourse unit that induces a discourse topic is a feeder which is typically at the start of a discourse, and it DOES NOT ANSWER ANY QUESTION because no context is available. A feeder may also exist when a new discourse topic arises.
Task: Given a text shown as a sequence of indexed sentences, 1. find the overall discourse topic of this text, and the topics under the discourse topic; 2. express topics with questions. 3. Each topic may dominate one or more sentences, and so show the boundaries of each topic (the start sentence id and end sentence id). Sentences under the same topic should be ADJACENT to each other. 4. Peform analysis recursively until QUESTION EACH SENTENCE ANSWERS is inferred. 5. QUESTIONS FOR EACH SENTENCE SHOULD NOT LEAK WORDS IN THE SENTENCE. ONLY ask questions based on the topic and the PRECEDING questions and sentences, but the question should be answerable by the following sentence. 6. organize the sentence questions, the topic-constituting questions, and the question for the discourse topic into a hierarchical structure. If a question appears in the text, use this explicit question, otherwise infer the question that each discourse unit answers. NO QUESTIONS FOR THE FEEDER
Input: wsj_0618
Output: "Q1": "parent": null, "answer": [1, 9], "question": "What changes are being proposed in the car dealership industry, and what are their implications?", "feeder": [1] , "Q1.1": "parent": "Q1", "answer": [1, 2], "question": "What specific advice is being given to car dealers regarding inventory management?" , "Q1.1.1": "parent": "Q1.1", "answer": [1], "question": null , "Q1.1.2": "parent": "Q1.1", "answer": [2], "question": "What is the specific advice?" , "Q1.2": "parent": "Q1", "answer": [3, 5], "question": "What are the current challenges in the car dealership industry that necessitate these changes?" , "Q1.2.1": "parent": "Q1.2", "answer": [3, 4], "question": "What led to the call for emergency action?" , "Q1.2.1.1": "parent": "Q1.2.1", "answer": [3], "question": "What has been the financial state of car dealers recently?" , "Q1.2.1.2": "parent": "Q1.2.1", "answer": [4], "question": "What is the general situation of the inventory?" , "Q1.2.2": "parent": "Q1.2", "answer": [5], "question": "What specific advice does Mr. Tonkin make, and what is the rationale?" , "Q1.3": "parent": "Q1", "answer": [6, 9], "question": "What are the reactions and potential consequences of these proposed changes?", "Q1.3.1": "parent": "Q1.3", "answer": [6, 7], "question": "How is the proposed plan being received?" , "Q1.3.1.1": "parent": "Q1.3.1", "answer": [6], "question": "What are the responses regarding the inventory reduction?" , "Q1.3.1.2": "parent": "Q1.3.1", "answer": [7], "question": "What actions have dealers taken?" , "Q1.3.2": "parent": "Q1.3", "answer": [8, 9], "question": "What criticisms are being raised against Mr. Tonkin's suggestions?" , "Q1.3.2.1": "parent": "Q1.3.2", "answer": [8], "question": How is the plan considered?" , "Q1.3.2.2": "parent": "Q1.3.2", "answer": [9], "question": "Why is it considered in this way?"
more examples: (wsj_0652, wsj_0613) example outputs:
Input: (a list of sentences) Output:

Table 9: The prompt template used in the experiments on QUD annotation using GPT-4.

F Detailed IAA Scores Between Human Annotator and GPT-4

Table 10 shows the comparison between human annotations and annotations by GPT-4 on short files. “Tree-h” indicates tree height when the set of trees is viewed as a single tree guided by a discourse topic.

Filenames	Human(Tree-h)	GPT(Tree-h)	κ
wsj_1381	3	3	0.7403
wsj_1985	3	3	0.4944
wsj_2313	4	3	0.3665
			mean: 0.5337

Table 10: Annotation results of GPT-4 on short files.

G Error Analysis of GPT-4’s Annotation on wsj_2313

In the list of sentences given below, the second sentence evaluates the typical approach to property buying, while the last sentence illustrates how the new approach functions. Consequently, the second sentence should follow the first one. However, GPT-4 groups the second and third sentences under the same topic and assigns the question ‘What is Kaufman & Broad’s usual operational approach?’ to the second sentence, which seems questionable.

- “Typically, developers option property, and then once they get the administrative approvals, they buy it,” said Mr. Karatz, adding that he believes the joint venture is the first of its kind.
- “We usually operate in that conservative manner.”
- By setting up the joint venture...

H Analysis of QUD Trees Proposed by Riester (2019)

An example is taken from Shahmohammadi et al. (2023)[10]). The English version, obtained through Google Translate, is provided in parentheses.

> Q11 Was ist das Thema der Episode? (What is the theme of this episode?)
>
> —Ich spreche mit Pavel Mayer, einem der vier Entwickler der Software über die Entstehungsgeschichte von Terravision, die Technikkultur der 90er, das neuartige User Interface “Earth Tracker”, wie es in der Folge zu der Auseinandersetzung mit Google kam (I speak with Pavel Mayer, one of the four developers of the software, about the history of Terravision, the technology culture of the 90s, the new user interface “Earth Tracker”, how the dispute with Google came about)
>
> —und wo der Film und die Realität übereinstimmen (and where the film and reality coincide)
>
> —und wo sie dramaturgisch bewusst nicht zusammenpassen. (and where they deliberately do not fit together dramatically.)

It can be seen that sub-sentential units jointly answer one QUD. However, there are also cases when questions are identified for such units. An example is shown below (same source as the above example):

> Q2 Wann wurde die Software entwickelt? (When was the software developed?)
>
> —Als die Software von ART+COM 1994 das Licht der Welt erblickte, (When ART+COM’s software was launched in 1994)
>
> —Q3 Wie waren die Reaktionen? (What were the reactions?)
>
> —war die Überraschung groß, (the surprise was great)
>
> —Q4 Warum? (Why?)
>
> —da man diese Fähigkeiten erst sehr viel später erwartet hatte. (because these capabilities were not expected until much later.)

Q2 and Q4 seem to involve discourse relations typically captured in RST and PDTB. However, the triggers are arguably only the words “wann” (“when”) and “da” (“because”). These questions address local information, involving no discourse-level understanding or inference of the relationship between the propositions of textual segments.

[10] https://github.com/mohamadi-sara20/rst-qud-comparison/tree/main/qud