Discourse Analysis via Questions and Answers: Parsing Dependency Structures of Questions Under Discussion

Wei-Jen Ko[1] **Yating Wu**[2] **Cutter Dalton**[3] **Dananjay Srinivas**[3]
Greg Durrett[1] **Junyi Jessy Li**[4]
[1] Computer Science, [2] Electrical and Computer Engineering, [4] Linguistics,
The University of Texas at Austin
[3] Linguistics, University of Colorado Boulder
wjko@outlook.com, yating.wu@utexas.edu, cutter.dalton@colorado.edu, dananjay.srinivas@gmail.com,
gdurrett@cs.utexas.edu, jessy@utexas.edu

Abstract

Automatic discourse processing is bottlenecked by data: current discourse formalisms pose highly demanding annotation tasks involving large taxonomies of discourse relations, making them inaccessible to lay annotators. This work instead adopts the linguistic framework of Questions Under Discussion (QUD) for discourse analysis and seeks to derive QUD structures automatically. QUD views each sentence as an answer to a question triggered in prior context; thus, we characterize relationships between sentences as free-form questions, in contrast to exhaustive fine-grained taxonomies. We develop the first-of-its-kind QUD parser that derives a dependency structure of questions over full documents, trained using a large, crowdsourced question-answering dataset DCQA (Ko et al., 2022). Human evaluation results show that QUD dependency parsing is possible for language models trained with this crowdsourced, generalizable annotation scheme. We illustrate how our QUD structure is distinct from RST trees, and demonstrate the utility of QUD analysis in the context of document simplification. Our findings show that QUD parsing is an appealing alternative for automatic discourse processing.

1 Introduction

Discourse structure characterizes how each sentence in a text relates to others to reflect the author's high level reasoning and communicative intent. Understanding discourse can be widely useful in applications such as text summarization (Hirao et al., 2013; Gerani et al., 2014; Durrett et al., 2016; Xu et al., 2020), classification (Bhatia et al., 2015; Ji and Smith, 2017), narrative understanding (Lee and Goldwasser, 2019), machine comprehension (Narasimhan and Barzilay, 2015), etc.

However, automatically inferring discourse structure is challenging which hinders wider application (Atwell et al., 2021). At its root lies the issue of data annotation: popular coherence formalisms like the Rhetorical Structure Theory (RST, Mann and Thompson (1988), Segmented Discourse Representation Theory (SDRT, Asher et al. (2003), and the Penn Discourse Treebank (PDTB, Prasad et al. (2008) require experts—typically linguists trained for the task—to reason through long documents over large relation taxonomies. These features, coupled with the difficulties of annotating full structures in the case of RST and SDRT, make the task inaccessible to lay annotators. The taxonomies differ across formalisms (Demberg et al., 2019), and their coverage and definitions are being actively researched and refined (Sanders et al., 1992; Taboada and Mann, 2006; Prasad et al., 2014).

In contrast, this work aims to derive discourse structures that fit into the linguistic framework of *Questions Under Discussion* (QUD) (Von Stutterheim and Klein, 1989; Van Kuppevelt, 1995), which neatly avoids reliance on a strict taxonomy. In QUD, "each sentence in discourse addresses a (often implicit) QUD either by answering it, or by bringing up another question that can help answering that QUD. The linguistic form and the interpretation of a sentence, in turn, may depend on the QUD it addresses" (Benz and Jasinskaja, 2017). Thus relationships between sentences can be characterized by free-form questions instead of pre-defined taxonomies. For instance, consider the following two sentences:

> **(S3)**: A route out of Sarajevo was expected to open later today — but only for international humanitarian agencies that already can use another route.
> **(S6)**: A four-month cease-fire agreement signed Dec. 31 made possible the medical evacuation and opening of the route into Sarajevo today.

Sentence 6 is the answer to a question from sentence 3: *"Why can they open a route?"*. The question-answer view is in line with recent work reformulating linguistic annotation as question answering (He et al., 2015; Pyatkin et al., 2020; Klein

Findings of the Association for Computational Linguistics: ACL 2023, pages 11181–11195

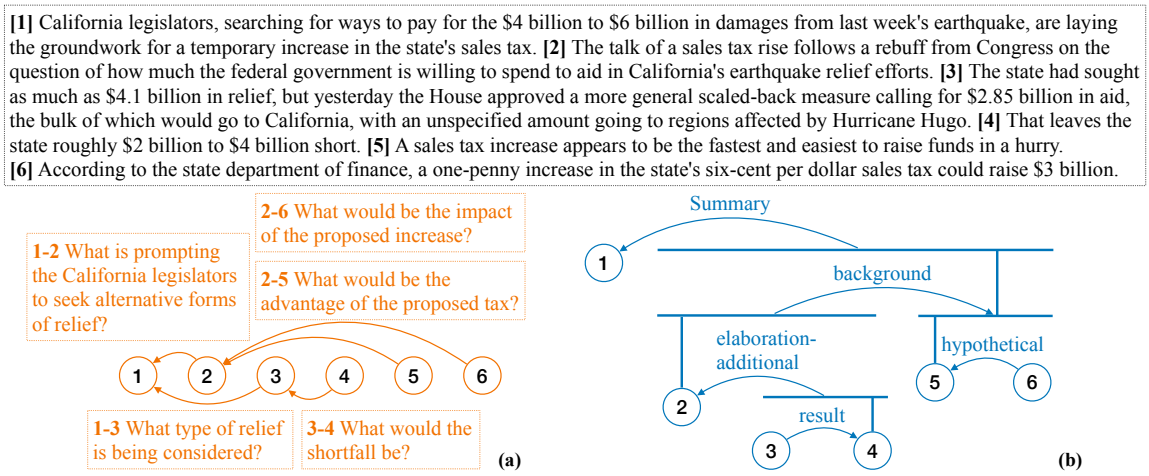

Figure 1: A snippet of a WSJ article from the intersecting subset of DCQA (Ko et al., 2022) and the RST Discourse Treebank (Carlson et al., 2001). (a) shows a QUD dependency structure derived from DCQA. Edges are defined by questions, connecting where the question arose from (the "anchor" sentence) and the sentence that answers the question. (b) shows the annotated RST tree above the sentence level.

et al., 2020), which reduces the bar for data collection and allows advancements in QA systems to be recruited (Aralikatte et al., 2021). Furthermore, QUD's reliance on natural language annotation aligns with large language models (e.g., GPT-3) using language as a universal "interface" across various tasks.

Despite the richness in theoretical research related to QUD, data-driven efforts are scarce; recent work has started corpora development under QUD (De Kuthy et al., 2018; Westera et al., 2020; Hesse et al., 2020), but these dedicated datasets are small and no computational models have yet been built to automatically derive QUD structures.

This work seeks to fill this gap, and presents the first-of-its-kind QUD parser. This parser takes a document as input and returns a question-labeled dependency structure over the sentences in the document, as depicted in Figure 1(a). For training, we use the intra-document question answering dataset DCQA (Ko et al., 2022); DCQA's annotation scheme is both compatible with QUD and easily crowdsourced, making QUD parsing a much less costly option than existing frameworks.

Each question in DCQA is considered to arise from an "anchor" sentence, and answered by another sentence later in the same article. In line with QUD, we consider each sentence as the answer to an implicit question from prior context (Hunter and Abrusán, 2015), in particular the anchor sentence. We view the anchor sentence as the parent node of the answer sentence, with the question describing the relation between the two; this results in a dependency tree structure.

Conveniently, a subset of DCQA overlaps with the RST Discourse Treebank (Carlson et al., 2001), allowing us to directly compare the two types of structures (Figure 1(b)). We show that the QUD trees are structurally distinct from RST trees. A close inspection of relation-question correspondence reveals that QUD's free-form questions are more fine-grained, and that their presence reduces annotator disagreement in selecting RST relations.

Trained on DCQA, our QUD parser consists of two models used in a pipeline. The first model predicts the anchor sentence for each (answer) sentence in the article; the second model performs question generation given the answer sentence and the predicted anchor sentence. Our comprehensive human evaluation shows that readers approve of 71.5% of the questions generated by our best model; among those, the answer sentence answers the generated question 78.8% of the time. Finally, we demonstrate the analytical value of QUD analysis in the context of news document simplification: the questions reveal how content is elaborated and reorganized in simplified texts.

In sum, this work marks the first step in QUD parsing; our largely positive human evaluation results show that this is a promising data-driven approach to discourse analysis with *open, crowdsourced* annotation that is so far infeasible to do at scale with other discourse frameworks. We release our models at `https://github.com/lingchensanwen/DCQA-QUD-parsing`.

2 Background and related work

Discourse frameworks Questions Under Discussion is a general framework with vast theoretical research especially in pragmatics, e.g., information structure (Roberts, 2012; Büring, 2003; Velleman and Beaver, 2016), presuppositions (Simons et al., 2010), and implicature (Hirschberg, 1985; Van Kuppevelt, 1996; Jasinskaja et al., 2017). Ginzburg et al. (1996) extended Stalnaker (1978)'s dynamic view of context to dialogue by integrating QUD with dialogue semantics, where the speakers are viewed as interactively posing and resolving queries. In QUD analysis of monologue, each sentence aims to answer a (mostly implicit) question triggered in prior context. Sometimes the questions form hierarchical relationships (stacks where larger questions have sub-questions, starting from the root question "*What is the way things are?*") (Büring, 2003; Roberts, 2004; De Kuthy et al., 2018; Riester, 2019). However, because of the inherent subjectivity among naturally elicited QUD questions (Westera et al., 2020; Ko et al., 2020), we leave question relationships for future work.

QUD and coherence structures are closely related. Prior theoretical work looked into the mapping of QUDs to discourse relations (Jasinskaja et al., 2008; Onea, 2016) or the integration of the two (Kuppevelt, 1996). Hunter and Abrusán (2015) and Riester (2019) studied structural correspondances between QUD stacks and SDRT specifically. Westera et al. (2020) showed that QUD could be a useful tool to quantitatively study the predictability of discourse relations (Garvey and Caramazza, 1974; Kehler et al., 2008; Bott and Solstad, 2014). In Pyatkin et al. (2020), discourse relation taxonomies were also converted to templatic questions, though not in the QUD context.

Traditionally, discourse "dependency parsing" refers to parsing the RST structure (Hirao et al., 2013; Bhatia et al., 2015; Morey et al., 2018). Since QUD structures are marked by free-form questions, the key aspect of "parsing" a QUD structure is thus question generation, yielding a very different task and type of structure than RST parsing. As we show in the paper, the two are complementary to each other and not comparable. This work focuses on automating and evaluating a QUD parser; we leave for future work to explore what types of structure is helpful in different downstream tasks.

The DCQA dataset Corpora specific for QUD are scarce. Existing work includes a handful of interviews and 40 German driving reports annotated with question stacks (De Kuthy et al., 2018; Hesse et al., 2020), as well as Westera et al. (2020)'s 6 TED talks annotated following Kehler and Rohde (2017)'s expectation-driven model (eliciting questions without seeing upcoming context). Ko et al. (2020)'s larger INQUISITIVE question dataset is annotated in a similar manner, but INQUISITIVE only provides questions for the first 5 sentences of an article, and they did not annotate answers.

This work in contrast repurposes the much larger **DCQA** dataset (Ko et al., 2022), consisting of more than 22K questions crowdsourced across 606 news articles. DCQA was proposed as a way to more reliably and efficiently collect data to train QA systems to answer high-level questions, specifically QUD questions in INQUISITIVE. Though not originally designed for QUD parsing, DCQA is suitable for our work because its annotation procedure follows the reactive model of processing that is standard in QUD analysis (Benz and Jasinskaja, 2017), where the questions are elicited after observing the upcoming context. Concretely, for each sentence in the article, the annotator writes a QUD such that the sentence is its answer, and identifies the "anchor" sentence in preceding context that the question arose from. Figure 1(a) shows questions asked when each of the sentences 2-6 are considered as answers, and their corresponding anchor sentences. As with other discourse parsers, ours is inevitably bound by its training data. However, DCQA's crowdsourcable paradigm makes future training much easier to scale up and generalize.

3 Questions vs. coherence relations

We first illustrate how questions capture intersentential relationships, compared with those in coherence structures. We utilize the relation *taxonomy* in RST for convenience, as in Section 5.3 we also compare the structure of our QUD dependency trees with that of RST.

Given each existing anchor-answer sentence pair across 7 DCQA documents, we asked two graduate students in Linguistics to select the most appropriate discourse relation between them (from the RST relation taxonomy (Carlson and Marcu, 2001)). Both students were first trained on the taxonomy using the RST annotation manual.

Analysis The frequency distribution of annotated RST relations that occurred ≥ 10 times (counting each annotator independently) is: *elaboration(200),*

cause(75), manner-means(69), background(64), explanation(55), comparison(33), condition(32), contrast(17), temporal(15), attribution(14). E.g.,

> ***[context]*** Early one Saturday in August 1992, South Floridians discovered they had 48 hours to brace for, or flee, ... one of the nation's most infamous hurricanes.
> ***[anchor]*** Oklahomans got all of 16 minutes before Monday's tornado.
> ***[QUD]*** How much time do people normally have to prepare for tornadoes?
> ***[answer]*** And that was more time than most past twisters have allowed.
>
> **RST label:** Comparison

Our analysis shows that the questions are often more fine-grained than RST relation labels; in the example below, the QUD describes what is being elaborated:

> ***[anchor]*** Crippled in space, the Kepler spacecraft's planet-hunting days are likely over.
> ***[QUD]*** What plans does NASA have for the damaged spacecraft?
> ***[answer]*** Engineers will try to bring the failed devices back into service, or find other ways to salvage the spacecraft.
>
> **RST label:** Elaboration-Additional

Agreeing on what is the most appropriate RST relation, as expected, is difficult with its large relation taxonomy: Krippendorff's α (with MASI distance to account for multiple selection) between the two annotators is 0.216, indicating only fair agreement (Artstein and Poesio, 2008). To study the effects of seeing the QUD, we further asked the annotators to find a relation *without* the question.[1] This led to a much lower, 0.158 α value. Thus the presence of the QUD could, in some cases, align divergent opinions, as in the following example:

> ***[context]*** For the past four years, the $600 million Kepler has been a prolific planet detector from its lonely orbit... ***[anchor]*** The project has been a stunning success, changing our view of the universe.
> ***[QUD]*** What knowledge did we have about solar systems before the project?
> ***[answer]*** Before Kepler, we knew little about other solar systems in the Milky Way galaxy.
>
> **RST labels with questions:** Background; Background
> **RST labels w/o questions:** Evidence; Circumstance

We also find that sometimes a question could be interpreted in terms of different RST relations:

> ***[anchor]*** According to a preliminary National Weather Service summary, Monday's tornado was a top-end EF5, with top winds of 200 to 210 miles per hour (mph), and was 1.3 miles wide.
> ***[QUD]*** How long did the tornado last?
> ***[answer]*** It was tracked on the ground for 50 minutes - an eternity for a tornado - and its damage zone is more than 17 miles wide.
>
> **RST labels that could work:** Evidence, Proportion, Elaboration-Additional, Manner

These findings indicate that while questions often relate to coherence relations, they are typically more specific and can also capture aspects from multiple relations. This supports Hunter and Abrusán (2015)'s skepticism about the correspondence of QUD and coherence structures, though they focused more on structural aspects of SDRT.

4 Deriving QUD dependency structures

Our task is to derive a QUD dependency structure over a document $D = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$ consisting of n sentences. A QUD tree $\mathbf{T} = ((a_1, \mathbf{q}_1), \ldots, (a_n, \mathbf{q_n}))$ can be expressed as a list of n tuples: each sentence has an associated anchor sentence a_i and a question labeling the edge to the anchor $\mathbf{q}_i$. To arrive at a dependency structure, we view the anchor sentence as the head of an edge, linking to the answer sentence via the question, as shown in Figure 1(a).

We set $a_1 = 0$ and $\mathbf{q}_1 = \emptyset$; the first sentence is always the root of the QUD dependency tree, so has no parent and no question labeling the edge. Each other $a_i \in \{1, 2, \ldots, i-1\}$ and $\mathbf{q}_i \in \Sigma^*$ for a vocabulary Σ. We note that $\mathbf{T}$ is analogous to a labeled dependency parse, except with questions $\mathbf{q}$ in place of typical discrete edge labels. Our parser is a discriminative model

$$P(\mathbf{T} \mid D) = \prod_{i=1}^{n} \left[P_a(a_i \mid D, i) P_q(\mathbf{q}_i \mid D, i, a_i) \right].$$

This formulation relies on models corresponding to two distinct subtasks. First, *anchor prediction* selects the most appropriate sentence in prior context to be the anchor sentence of the generated question using a model $P(a_i \mid D, i)$. Second, *question generation* given the current (answer) sentence, its anchor, and the document context uses a model $P(\mathbf{q}_i \mid D, i, a_i)$.

We do not impose projectivity constraints or other structural constraints beyond anchors needing to occur before their children. Therefore, inference can proceed with independent prediction for each sentence.[2] We now proceed to describe the models

[1] We paced 3 months between annotation with and without the question to minimize memorization effects.

[2] We make a further simplifying assumption by doing greedy prediction of each a_i before generating $\mathbf{q}$. We sample $\mathbf{q}$ using nucleus sampling and do not rely on the question probabilities to be informative about whether the structure itself is well-formed.

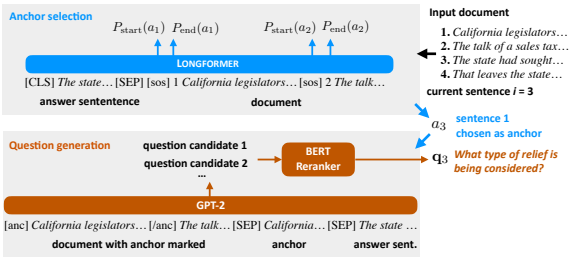

Figure 2: Breakdown of the two modules used in the parser: anchor selection using a Longformer QA model to select an anchor index and question generation conditioned on the selected anchor and the answer, including a BERT reranking phase.

for P_q and P_a that constitute the parser.

4.1 Anchor prediction

The anchor prediction model P_a considers the given sentence $\mathbf{s}_i$ and reasons through prior article context to find the most likely sentence where a QUD can be generated, such that $\mathbf{s}_i$ is the answer. Since this task involves long document contexts, we use the Longformer model (`longformer-base-4096`) (Beltagy et al., 2020), shown to improve both time efficiency and performance on a range of tasks with long contexts.

We adopt the standard setup of BERT for question answering (Devlin et al., 2019) and model $P(a_i)$ as a product of start and end distributions. For the input, we concatenate the answer sentence and the article as a single long sequence, separated by delimiters: `[CLS] [answer sentence] [SEP] [document]`. Following Ko et al. (2022), we add two tokens: the start of sentence token `[sos]` and the sentence ID, before every sentence in the article. We train the model to predict the span of the two added tokens in front of the anchor sentence.

We modify the HuggingFace (Wolf et al., 2020) codebase for our experiments. We use the Adam (Kingma and Ba, 2015) optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$ and learning rate 5e-5. The model is trained for 25000 steps using batch size 4. We use the same article split for training, validation and testing as in DCQA, and the parameters are tuned on the validation set.

4.2 Question generation

Our question generator $P_q(\mathbf{q}_i \mid D, i, a_i)$ takes in the answer sentence $\mathbf{s}_i$ indexed by i, the anchor sentence at a_i, and the article D, and aims to generate an appropriate QUD. We fine-tune GPT-2 (Radford et al., 2019) for this purpose; Ko et al. (2020) showed that GPT-2 generates open-ended, high-level questions with good quality. To fine-tune this model, each input instance is a concatenation of four parts, separated by delimiters: (1) $\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_{i-1}$, with the start and end of the anchor sentence marked by special tokens; (2) the anchor sentence; (3) $\mathbf{s}_i$; (4) the question.

Inference During inference, we feed in (1)—(3) and sample from the model to generate the question. By default, we use nucleus sampling (Holtzman et al., 2020) with $p = 0.9$. To improve the consistency of questions with the anchor or answer sentences, we use an additional **reranking step**.

Our reranker is a BERT binary classification model formatted as `[CLS] [question] [SEP] [anchor sentence] [answer sentence]`. Positive examples consist of annotated questions, anchor, and answer sentences in the DCQA training set; we synthetically generate negative examples by replacing the anchor or answer sentences with others in the same article. Training is detailed in Appendix B. To rerank, we sample 10 questions from the generation model, and choose the question with the highest posterior from the reranker.

Reducing question specificity We found that questions generated by the above model often copy parts of the answer sentence, including information that is introduced for the first time in the answer sentence. For example, in Figure 1, Hurricane Hugo is first mentioned in sentence 3. The model might ask "*What type of relief is going to California and regions affected by Hurricane Hugo?*" This makes the question prone to "foresee" details that are unlikely to be inferred from previous context, violating QUD principles. We observe that these unwanted details often pertain to specific entities. To this end, in the answer sentence, we replace each token that belongs to a named entity with its entity type before feeding into the GPT-2 model.[3]

5 Evaluation and analysis

Since QUD parsing features an open-ended generation component, we need new evaluation methodology compared to standard discourse parsing. We focus on two main factors: (1) whether the generated question is plausible at the predicted anchor point; (2) whether the question is actually answered by the answer sentence.

[3]We use the `bert-base-NER` model trained on the CoNLL-2003 NER dataset (Sang et al., 2003)

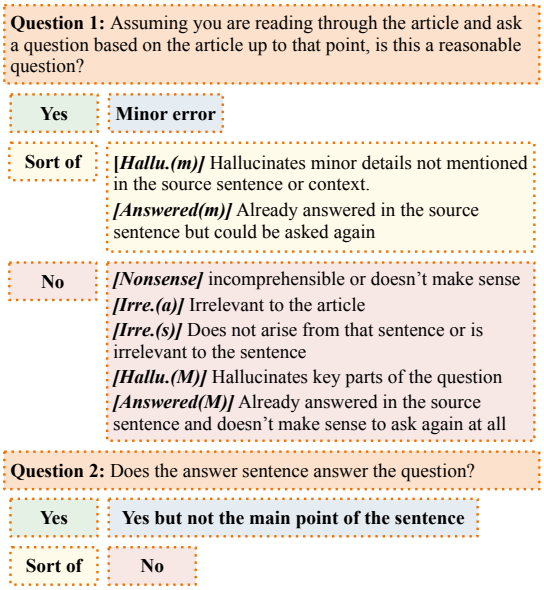

Figure 3: Evaluation schema.

In QUD annotation and DCQA itself (Westera et al., 2020; Ko et al., 2020, 2022), it is often the case that multiple questions can be asked even given the same anchor and/or answer sentences. The evaluation of QUD validity thus involves complex reasoning performed jointly among (long) context, the anchor, the answer sentence, and the generated question itself. For these reasons, we rely on human evaluation, and leave the development of automatic evaluation metrics for future work.[4]

5.1 Human Evaluation Setup

Our evaluation task shows human judges the full article, the anchor and answer sentences, and the generated question. We then ask them to judge the quality of the generated QUD using a hierarchical schema shown in Figure 3. The criteria in our evaluation overlap with De Kuthy et al. (2018)'s human annotation guidelines, while specifically accommodating typical errors observed from machine-generated outputs.

Question 1 (Q1) assesses how reasonable the question is given context prior to and including the anchor sentence. The judges have four graded options: (1) *yes* for perfectly fine questions; (2) *minor error* for questions that contain minor typos or grammatical errors that do not impact its overall good quality; (3) *sort of* for questions with non-negligible though not catastrophic errors; and (4) *no* for questions that are not acceptable. (3) and (4) both contain subcategories representative for a sample of questions we closely inspected a priori.

Question 2 (Q2) assesses whether the question is answered by the targeted answer sentence, also with four graded options: (1) *yes* where the targeted answer sentence is clearly an answer to the generated question; (2) *yes but not the main point* where the answer is not the at-issue content of the answer sentence. Such cases violate Grice's principle of quantity (Grice, 1975) and QUD's principle that answers should be the at-issue content of the sentence (Simons et al., 2010). (3) *sort of* where the answer sentence is relevant to the question but it is questionable whether it actually addresses it; and (4) *no* where the generated question is clearly not addressed by the answer sentence. Annotators are allowed to skip Q2 if the generated question from Q1 is of lower quality.

5.2 Results

We recruited 3 workers from Mechanical Turk as judges who have an established relationship with our lab, and are experienced with tasks involving long documents. They are compensated above $10 per hour. We annotate 380 questions from 20 articles from the DCQA test set. Inter-annotator agreement is reported in Appendix A.

Q1 results As seen in Table 1, for our full model, 71.5% of responses are "yes"es, showing that most of the generated questions are of good quality. Without reranking, there are 4.8% fewer "yes" responses; there are more questions that do not rise from the anchor sentence, showing the effectiveness of our reranker. Further removing NER masking results in a substantial drop of 11.9% of good questions. There are also more questions hallucinating details and/or irrelevant to the anchor sentence.

Q2 results Since Question 2 may not make sense when the generated question is of low quality, we show the results of Q2 on a subset of questions where all three workers answered "yes" or "minor error" for Q1 (see Table 2). Of those questions, annotators chose "yes" 78.8% of the time, showing that a majority of good-quality questions are actually answered in the answer sentence and represents anchor-answer sentence relationships. Our full model has better performance than the two

[4]Existing automatic measures for open-ended tasks are known to correlate poorly with human judgments (Howcroft et al., 2020; Celikyilmaz et al., 2020); additionally, whether the answer sentence actually answers the question is a key aspect to validate QUD. But as shown in Ko et al. (2022), poor performance bars existing QA models from being used to evaluate QUD parsing. Appendix C discusses anchor prediction "accuracies" against human-annotated anchors in DCQA.

System	Yes	Minor error	Sort of		No				
			Hallu.(m)	Ans.(m)	Nonsense	Irre.(a)	Irre.(s)	Hallu.(M)	Ans.(M)
Full	71.5	4.2	7.1	4.0	6.4	0.2	3.0	2.4	1.2
-Reranking	66.7	3.4	8.4	4.5	6.3	0.2	7.8	1.7	1.0
-NER	54.8	2.8	10.7	4.2	6.2	0.6	16.9	2.9	1.0

Table 1: Human evaluation results for Question 1.

System	Yes	Not main point	Sort of	No
Full	78.8	3.1	10.5	7.6
-Reranking	71.8	1.8	14.1	12.3
-NER	76.7	2.8	11.0	9.4

Table 2: Human evaluation results for Question 2.

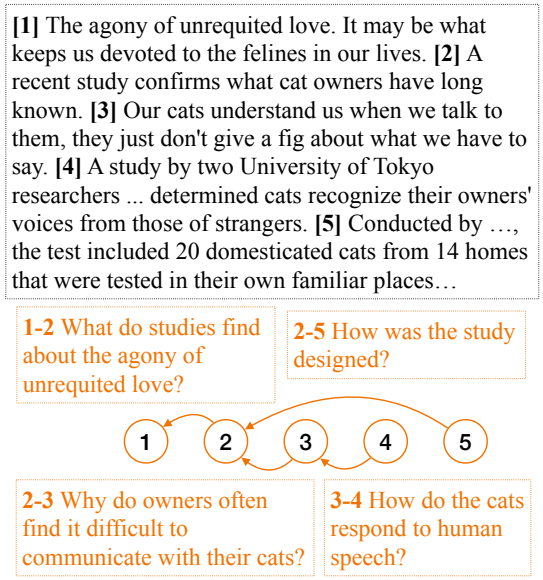

Figure 4: Example of model-generated QUD structure.

ablations, showing the effectivenes of reranking. Further, since masking NER removes some of the information from the answer sentence, the percentage of "yes"es is slightly lower after masking.

These results show that most of the time, our full system is able to generate questions that are good in terms of linguistic form and are also reasonable QUD questions given prior context. Most of these good questions are clearly answered in the answer sentence, i.e., they legit questions under the reactive model of mental processing. These results indicate a strong QUD parser with a large portion of valid QUD links. In Figure 4 and Appendix D, we visualize output examples.

5.3 Characterizing tree structures

We further characterize annotated and parsed QUD trees; we also contrast QUD trees with RST, using the intersection of DCQA and RST-DT (Carlson et al., 2001). We follow Hirao et al. (2013) to convert RST constituency trees to dependency trees using nuclearity information. Since the leaves of QUD trees are sentences, we also treat sentences as the smallest discourse units for RST.

We report the six metrics following (Ferracane et al., 2019): 1) tree **height**; 2) **normalized arc length**: the average number of sentences between edges, divided by the number of sentence n in the article; 3) **proportion of leaf nodes**: the number of leaf nodes divided by n; 4) **average depth** of every node in the tree; 5) **right branch**: the number of nodes whose parent is the immediate preceding sentence in the article, divided by n; 6) **attachment score**: count of sentences whose parent node is the same sentence among the two types of trees, divided by n, the total number of sentences. This captures the similarity of the two types of trees.

Compared with annotated QUD trees, machine generated ones are slightly deeper and more right-branching (Table 3). The normalized arc lengths indicate that our model is not merely finding the immediately preceding sentence as the anchor, although human annotated trees tend to have slightly longer arc lengths. Machine-derived trees have a lower gap degree (Yadav et al., 2019) (13.2 on average on the validation set), compared to annotated ones (15.1 on average).

5.4 QUD vs. RST

Compared with RST (Table 3), QUD trees have longer arc lengths, showing that they more frequently represent relations between more distant sentence pairs. The tree height and average node depth of DCQA trees are larger than those of RST.

While nuclearity in RST is able to provide a hierarchical view of the text that has been used in NLP tasks, it comes with a highly contested (Wolf and Gibson, 2005; Taboada and Mann, 2006) strong compositionality assumption that "whenever two large text spans are connected through a rhetorical relation, that rhetorical relation holds also between the most important parts of the constituent spans" (Marcu, 1996). Marcu (1998) showed that this assumption renders the derived structure alone

data	tree type	height	norm. arc len.	prop. of leaf	avg. depth	right branch	att. score
RST ∩ DCQA	RST-dep	5.86	0.12	0.53	3.49	0.40	0.30
RST ∩ DCQA	DCQA-human	6.72	0.21	0.48	3.88	0.45	
DCQA (val)	DCQA-human	6.04	0.29	0.50	3.57	0.39	0.47
DCQA (val)	DCQA-model	6.76	0.22	0.43	3.85	0.52	

Table 3: Statistics of discourse dependency trees, on the intersecting documents of RST-DT and DCQA (upper portion) and the DCQA validation set (lower portion).

insufficient in text summarization. In contrast, the QUD framework does not make such an assumption since it does not have the RST notion of nuclearity. During left-to-right reading, QUD describes how each sentence resolves an implicit question posed in prior context, so QUD dependencies derived in this work are always rooted in the first sentence and "parentage" does not necessarily entail salience. Combined with observations from Section 3, we conclude that RST and QUD are complementary frameworks capturing different types of structure.

6 Case study: document simplification

We demonstrate the analytical value of QUD analysis in the context of document simplification. We use the Newsela dataset (Xu et al., 2015), where news articles are professionally simplified across several grade levels; a subset of Newsela (of the highest reading level) is present in DCQA. Note that most research in text simplification focus on the sentence level (Alva-Manchego et al., 2020); we hope to inform document-level approaches.

We sample 6 articles from the DCQA Newsela subset. For each of these, 3 linguistics undergraduates (not authors of this paper) doubly annotated their corresponding middle and elementary school levels with QUD structures for the first 20 sentences following DCQA's paradigm. This amounts to ∼720 questions in total. Figure 5 shows a snippet of our analysis from two reading levels of the same article.

We run and evaluate our parser on the articles of the second reading level. Using the schema in Figure 3, Question 1 is *yes* for 60.2% of the time, and Question 2 is *yes* 75.2% of the time. This shows that while the parser is still capable of generating reasonable questions, the performance degrades compared to testing on the highest level. This is likely due to clear stylistic, organizational, and vocabulary difference for simplified texts; for this reason, we resort to using annotated QUDs to illustrate idealized results for this analysis.

Analysis The simplified articles, which mostly align with the original versions at the beginning, tend to contain significant reorganization of content especially later in the text. Nonetheless, we found that 62.2% of the questions had a similar question on another reading level, reflecting that QUDs frequently stay invariant despite these differences. For example, in Figure 5, the content of sentence 8 (level 2) is covered in sentence 2 (level 1), yet in both cases the question "*Why is the case important*" is used to link these sentences. Similarly, questions `q2-6` (level 2) and `q2-8` (level 1), as well as questions `q6-7` (level 2) and `q8-10` (level 1) reflect the same QUD.

Often, articles from higher reading levels presupposes certain knowledge that gets **elaborated** or explained during simplification (Srikanth and Li, 2021). QUD analysis informs how content should be elaborated: in Figure 5(a), the level 1 article defined the concept of amendment (question `q8-9`), absent in level 2.

Sentence splitting as a frequent operation (Petersen and Ostendorf, 2007; Zhu et al., 2010; Alva-Manchego et al., 2020) could also be explained by questions, as in the case of `q8-11` in level 1, which provides a rationale as to why sentence 8 in level 2 is split (into sentences 2 and 11 in level 1). Note that this explanation is rooted *outside* of content conveyed by the sentence that was split.

Finally, editors also **omit** difficult content (Petersen and Ostendorf, 2007; Zhong et al., 2020), as in Figure 5(b): sentence 1 in level 2 is not present in the level 1 simplification (due to less salience and the reference to the "selfie generation" which goes beyond the targeted reading level). Level 2 thus contains the extra QUD: `q1-2`.

In sum, QUD analysis reveals how elaborated or omitted content fit into the larger context during simplification, potentially aiding future document-level simplification systems by providing intermediate rationales.

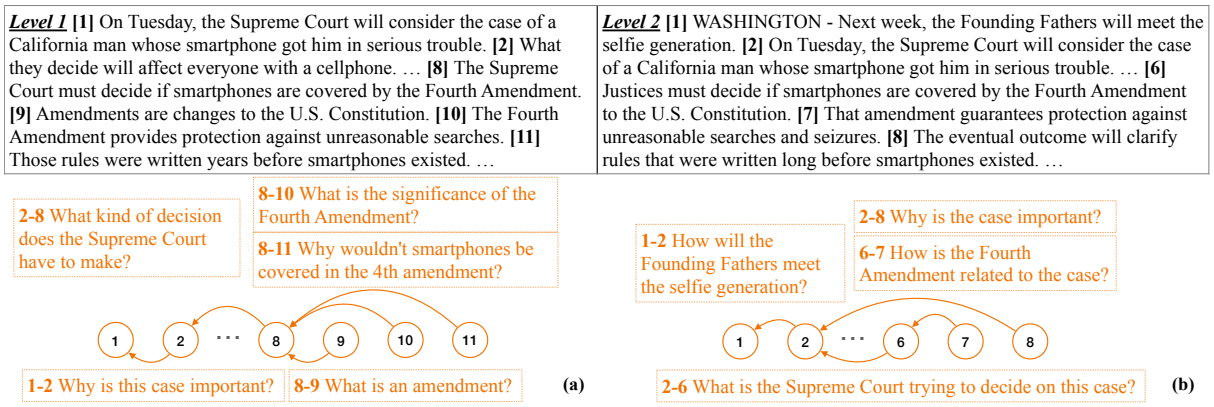

Figure 5: Snippet of a Newsela article set with two reading levels.

7 Conclusion

This work presents the first QUD (Questions Under Discussion) parser for discourse analysis. We derive dependency structures of QUD, viewing each sentence as an answer to a QUD triggered in an anchor sentence in prior context. This paradigm avoids costly annotation of coherence structures; rather, our parser can be trained on the crowdsourced dataset DCQA. We show strong parser performance with comprehensive human evaluation. We further demonstrate the richness of QUD analysis in document simplification.

8 Limitations

While our work is consistent with the key aspects of Questions Under Discussion, we do not attempt to take into account all aspects of this broad framework. Most notably, we do not model relationship between questions (or question stacks), as mentioned in Section 2. While such relationships are potentially useful, with question stacks, the annotation task becomes much more expensive; currently, no existing dataset is available to train parsers in this fashion. We applaud the development of tools such as TreeAnno (De Kuthy et al., 2018) to aid annotation. Additionally, because questions are open-ended, they are inherently subjective, which adds substantial challenge to modeling and evaluating stacks. Constrained by DCQA's setup, we also do not explicitly model QUD with multi-sentence answers, and leave this for future work.

The subjectivity of QUD analysis also means that there is no single "right" structure. This is in contrast to coherence structures that more rigorously define their structures and relation taxonomies (multiple analyses still exist in those structures, but to a lesser degree). Nonetheless, we showed in Section 6 that consistency is still present despite documents being reworded and restructured during simplification.

To evaluate our parser, we developed a human evaluation scheme. As mentioned in Section 5, automatic evaluation of QUD structure contains both a generation and a question-answering component. However, human evaluation is costly; future work looking into the development of automatic evaluation measures can be extremely valuable.

Acknowledgments

We thank Kathryn Kazanas, Keziah Reina, and Anna Alvis for their contributions on text simplification analysis. We thank David Beaver for helpful discussions and comments. This research is partially supported by NSF grants IIS-2145479, IIS-2107524. We acknowledge the Texas Advanced Computing Center (TACC)[5] at UT Austin for many of the results within this paper.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Rahul Aralikatte, Matthew Lamm, Daniel Hardt, and Anders Søgaard. 2021. Ellipsis resolution as question answering: An evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 810–817.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

[5] https://www.tacc.utexas.edu

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

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse. *Discourse Processes*, 54(3):177–186.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218.

Oliver Bott and Torgrim Solstad. 2014. From verbs to discourse: A novel account of implicit causality. In *Psycholinguistic approaches to meaning and understanding across languages*, pages 213–251.

Daniel Büring. 2003. On d-trees, beans, and b-accents. *Linguistics and philosophy*, 26(5):511–545.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54(2001):56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *SIGdial Workshop on Discourse and Dialogue*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

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*.

Vera Demberg, Merel CJ Scholman, and Fatemeh Torabi Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue & Discourse*, 10(1):87–135.

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.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2019. Evaluating discourse in structured text representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 646–653.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic inquiry*, 5(3):459–464.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1602–1613.

Jonathan Ginzburg et al. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 643–653.

Christoph Hesse, Anton Benz, Maurice Langner, Felix Theodor, and Ralf Klabunde. 2020. Annotating quds for generating pragmatically rich texts. In *Proceedings of the Workshop on Discourse Theories for Text Planning*, pages 10–16.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1515–1520.

Julia Linn Bell Hirschberg. 1985. *A theory of scalar implicature*. University of Pennsylvania.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration . In *International Conference on Learning Representations*.

David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Julie Hunter and Márta Abrusán. 2015. Rhetorical structure and quds. In *JSAI International Symposium on Artificial Intelligence*, pages 41–57.

Katja Jasinskaja, Fabienne Salfner, and Constantin Freitag. 2017. Discourse-level implicature: A case for qud. *Discourse Processes*, 54(3):239–258.

Katja Jasinskaja, Henk Zeevat, et al. 2008. Explaining additive, adversative and contrast marking in russian and english. *Revue de Sémantique et Pragmatique*, 24(1):65–91.

Yangfeng Ji and Noah A Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 996–1005.

Andrew Kehler, Laura Kertz, Hannah Rohde, and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of semantics*, 25(1):1–44.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. Qanom: Question-answer driven srl for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083.

Wei-Jen Ko, Te-yuan Chen, Yiyan Huang, Greg Durrett, and Junyi Jessy Li. 2020. Inquisitive question generation for high level text comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6544–6555.

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.

Jan van Kuppevelt. 1996. Directionality in discourse: Prominence differences in subordination relations1. *Journal of semantics*, 13(4):363–395.

I-Ta Lee and Dan Goldwasser. 2019. Multi-relational script learning for discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4214–4226.

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.

Daniel Marcu. 1996. Building up rhetorical structure trees. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1069–1074.

Daniel Marcu. 1998. To build text summaries of high quality, nuclearity is not sufficient. In *Working Notes of the AAAI-98 Spring Symposium on Intelligent Text Summarization*, pages 1–8.

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1253–1262.

Edgar Onea. 2016. *Potential questions at the semantics-pragmatics interface*. Brill.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Language Resources and Evaluation Conference*.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

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*, pages 2804–2819.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

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

Craige Roberts. 2004. Context in dynamic interpretation. *The handbook of pragmatics*, 197:220.

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

Ted JM Sanders, Wilbert PM Spooren, and Leo GM Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse processes*, 15(1):1–35.

Tjong Kim Sang, Erik F., and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Semantics and linguistic theory*, volume 20, pages 309–327.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137.

Robert C Stalnaker. 1978. Assertion. In *Pragmatics*, pages 315–332. Brill.

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

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

Jan Van Kuppevelt. 1996. Inferring from topics: Scalar implicatures as topic-dependent inferences. *Linguistics and philosophy*, pages 393–443.

Leah Velleman and David Beaver. 2016. Question-based models of information structure. In *The Oxford handbook of information structure*.

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.

Matthijs Westera, Laia Mayol, and Hannah Rohde. 2020. TED-Q: TED talks and the questions they evoke. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1118–1127.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Himanshu Yadav, Samar Husain, and Richard Futrell. 2019. Are formal restrictions on crossing dependencies epiphenominal? In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 2–12.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9709–9716.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1353–1361.

A Inter-annotator agreement for human judgments

For **Question 1**, the three annotators all agree on 54% of the fine grain labels, and there is a majority on 93% of questions on fine grained labels. Krippendorff's alpha is 0.366 for "yes" vs. others, 0.319 for the 4 coarse categories, and 0.317 for all labels at the most fine-grained label. For **Question 2**, the three annotators all agree on 60% of the fine grain labels, and there is a majority on 93% of questions on fine grained labels. Krippendorff's alpha is 0.376 for "yes" vs. others, and 0.297 for the 4 categories.

All the alpha values above indicate "fair" agreement (Artstein and Poesio, 2008). One reason for this is a clear majority of "yes" labels for both questions; nonetheless these values indicate a certain degree of subjectivity in the tasks.

B Reranker details

To train the reranker, we use questions in the DCQA training set as positive examples, and swap the answer or the anchor sentence with *every* other sentence from the same article to create negative examples. This resulted in a training set of 709,532 instances. We fine-tune the BERT model on this data for 3 epochs using learning rate 2e-5 and batch size 32, trained using binary cross entropy loss.

On the DCQA validation set, among about 37 options generated from the same question, the ranks of the correct response predicted by the model is on the 14% percentile in average.

C Anchor prediction

We also report the accuracy of the predicted anchor sentences for the first part of our pipeline model (i.e., before the questions get generated). Note that this is a partial notion of accuracy for analysis purposes, since it is natural for different questions to be triggered form different sentences (and sometimes perfectly fine for the same question to come from different sentences) (Ko et al., 2022). On the validation and test set of the DCQA dataset, and the agreement between the model and human on 46.8% of the instances (the annotations of different annotators are treated as separate instances). This is the same as DCQA's statistics between two human annotators.

D Example model outputs

We show an additional snippet of example model output:

> **Context: [9]** In 1971, Sierra Nevada bighorns were one of the first animals listed as threatened under the California Endangered Species Act. **[10]** In 2000, the federal government added the bighorns to its endangered lists. **[11]** 'There was a lot of concern about extinction,' says state biologist Tom Stephenson, the recovery project leader. **[12]** 'But with some good fortune and the combination of the right recovery efforts, it's gone as well as anybody could've imagined'. **[13]** Teams of biologists and volunteers in 2000 began their research, and in 2007 started reintroducing the Sierra Nevada bighorn by dispersing them into herds along the Sierra's crest. **[14]** The agencies designated 16 areas for the bighorns with the initial goal of repopulating 12 of them.
>
> **9-10**: What happened after that?
>
> **10-11**: What was the opinion of those involved in the recovery project?
>
> **9-12**: What happened to the bighorns?
>
> **12-13**: How did recovery efforts eventually go?
>
> **13-14**: How many areas were to be re-population based on the initial work?

E Compute

For all models in this work, we used 2 compute nodes each consisting of 3x NVIDIA A100 GPUs.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

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

B ☑ Did you use or create scientific artifacts?

Throughout the paper

☑ B1. Did you cite the creators of artifacts you used?
Throughout the paper

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix E

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Throughout the paper

☑ 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 5

C ☑ Did you run computational experiments?

Sections 3-7

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4, Appendix F

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☐ 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?
Not applicable. Left blank.

☐ 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.)?
Not applicable. Left blank.

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

Section 5

☑ 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.?
Section 5

☑ 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 5.2

☐ 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?
Not applicable. Data collection does not involve human subjects or demographic information

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. This is IRB-excempt since data collection does not involve human subjects or demographic information

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Data collection does not involve human subjects or demographic information