Discursive Socratic Questioning: (Unsupervised) Interpreting Neural Language Models for Discourse Understanding

Anonymous ACL submission

Abstract

Do neural language models (NLMs) understand the discourse they are processing? Traditional interpretation methods that address this question require pre-annotated explanations, which defeats the purpose of unsupervised explanation. We propose unsupervised Discursive Socratic Questioning (**DISQ**), a two-step interpretative measure.

DISQ first generates Socratic-style questions about the discourse and then queries NLMs about these questions. A model's understanding is measured by its responses to these questions. We apply DISQ to examine two fundamental discourse phenomena, namely discourse relation and discourse coherence. We find NLMs demonstrate non-trivial capacities without being trained on any discourse data: Q&A pairs in DISQ are shown to be evidence for discourse relation and cohesive devices for discourse coherence. DISQ brings initial evidence that NLMs understand discourse through reasoning. We find larger models perform better, but contradictions and hallucinations are still problems. We recommend DISQ as a universal diagnostic for discursive NLMs and using its output for self-supervision.

1 Introduction

Neural language models (NLMs) are criticized as not understanding text in the manner that humans do, in a logical and reliable way (Bender and Koller, 2020; Zhang et al., 2022; Tan et al., 2021). We study whether NLMs understand discourse, a fundamental linguistic subject concerning the organization of sentences. To understand discourse, humans usually identify key spans across multiple sentences and infer logical connections among them (Halliday, 1976; Camburu et al., 2018; Lei et al., 2018). We believe that the discourse community has largely ignored such intuitions, favoring the development of complex black-box models, where NLMs are leveraged as backbones (Liu et al., 2020). While achieving good performance, such black-box models lack interpretability and offer little evidence to trust their decisions. We believe it is imperative to examine the root cause: whether and how NLMs capture the linguistic properties of discourse function.

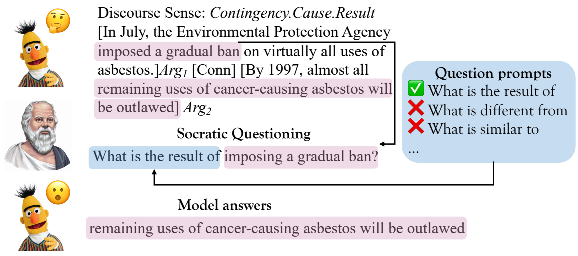


Figure 1: Discursive Socratic Questioning (DISQ) performs unsupervised interpretation. Step 1: Socratic-style questions are automatically generated by combining spans in discourse and question prompts. Step 2: A model answers these questions. Its output to the questions is used as a proxy for its understanding.

Popular interpretation methods like linguistic probing (Tenney et al., 2019) and behavior analyses have been shown as plausible methods (Belinkov et al., 2020; Choudhury et al., 2022). However, they have a major shortcoming: they require supervision. Additional annotations are required to train a model to predict linguistic structures or to generate explanations, which makes these methods difficult to apply to new tasks. We explore unsupervised interpretation as a novel alternative. In *Socratic Questioning* (named after the philosopher Socrates), a teacher raises thoughtful questioning to students to enable them to examine their ideas rigorously. At the end of the questioning, the students can determine the validity of the idea and discover any flaws and contradictions (Padesky, 1993).

We instantiate this idea for discourse understanding in the form of the Discursive Socratic Questioning procedure (**DISQ**; Figure 1). We enable NLMs to self-interrogate its understanding through Socratic-style questions. The premise is simple: if a text has a *Contingency* discourse relation, there must be a *cause* and a *result* (in purple). So if a NLM models the discourse appropriately, it must be able to also answer "what is the result of" question correctly, and must abstain from answering irrelevant questions. A battery of Socratic-style questions is created by combining question prompts (in blue) and text spans taken from the discourse. The model is self-interrogated by all questions, and we use the model's behavior as a proxy for its discourse understanding. We use a pre-determined set of question prompts (*c.f.* §2.1) to generate our questions, such that no additional supervision aside from discourse annotations are needed. We only need a text with discourse annotated or where the discourse is explicitly indicated (e.g., explicit discourse markers).

Through DISQ, we provide evidence that NLMs appropriately model discourse by reasoning over text as a set of key spans and inferring relationships among them, similar to how humans process discourse. (1) **DISQ identifies evidence for discourse relation.** We find Q&A in DISQ exhibits a strong association between question prompts and all four first-level discourse relations in the PDTB (Prasad et al., 2008). We also find that explicit discourse connectives boost the performance of the Socratic questioning. (2) **DISQ identifies cohesive devices for discourse coherence**: We consider Q&A pairs extracted by DISQ as cohesive devices. Simply aggregating them leads to a decent human correlation in SummEval dataset.

We present the first study using questioning for unsupervised model interpretation, with a focus on discourse understanding. Although in this study, we only examine standard English corpora, our DISQ reveals NLMs' non-trivial discourse modeling. We recommend that DISQ be used to serve as a universal diagnostic for NLM's representation of discourse, complementary to dataset benchmarking (Chen et al., 2019). Like Socrates did with his students, DISQ also diagnoses what a model *knows and does not know*. We observe that interesting patterns emerge, such as symmetry, self-contradiction, and hallucination in DISQ's output. We recommend two usability tests that utilize DISQ to help models diagnose their trustworthiness and use DISQ's output as self-supervision signals for future discursive NLMs. [1]

[1] We will release our codebase upon acceptance.

2 Discursive Socratic Questioning

What Counts as Discourse Understanding? Organized text makes sense as textual elements link the discourse together. Such linking elements are referred to as cohesive devices (Halliday, 1976). Concretely, given two discourse arguments Arg_1 and Arg_2 participating in a discourse relation R, two contiguous spans $s_1 \in Arg_1$ and $s_2 \in Arg_2$ link the two arguments into a coherent discourse with a semantic relation r. We define (s_1, s_2, r) as evidence for understanding the discourse relation. We argue that a model must be able to identify them for us to claim that it understands the discourse.

Discourse relation: Contingency.Cause.Result [In July, the Environmental Protection Agency [imposed a gradual ban]$_{s_1}$ on virtually all uses of asbestos.]$_{Arg_1}$ [By 1997, almost all [remaining uses of cancer-causing asbestos will be outlawed]$_{s_2}$]$_{Arg_2}$
Question: What is the result of imposing a ban? **Answer:** remaining uses of cancer-causing asbestos will be outlawed.

Table 1: Formalizing discourse understanding as question answering (QA). A *cause/result* relation between s_1 and s_2 is realized through QA.

Defining a Proxy for Discourse Understanding: We approach the notion of understanding through question answering (QA). We interrogate the model with a set of questions concerning different semantic relations and text spans. If a model is said to understand, it must answer questions in a manner consistent with the discourse relation.

As illustrated in Table 1, we believe NLMs must infer the *cause/result* relation r between "the ban" ($s_1 \in Arg_1$) and "remaining use of cancer-causing asbestos will be outlawed" ($s_2 \in Arg_2$) to understand the *contingency* discourse relation R. When querying about s_1, the model should extract s_2 as the answer with only *"what is the result"* prompt. It should not respond to irrelevant questions like *"what is different from"* since there is no such semantic relation to form a cohesive tie in the given discourse. An ideal model will extract all evidence triplets with only correct questions, and abstain from answering irrelevant questions.

Our approach is a generalized extension to Halliday (1976)'s theory. Halliday defined a taxonomy of cohesive devices, including reference, ellipsis, and lexical cohesion. These devices describe a limited set of specific text cohesion devices with constrained definitions. DISQ extends this computationally to encompass a more inclusive notion of cohesion among arbitrary spans in text and a larger relation space characterized by Socratic questioning (detailed discussion is in Appendix B).

2.1 Questioning and Answering

We operationalize DISQ with extractive QA to discover evidence (s_1, s_2, r) for understanding:

$$s_2 = \texttt{QA}(c = Arg_1 + Conn + Arg_2, q = p + s_1) \quad (1)$$

The model seeks an answer in the opposing discourse argument. The semantic relation r between s_1 and s_2 is determined by the question prompt p. Without loss of generality, if the question q is generated from $s_1 \in Arg_1$, then the answer must come from $s_2 \in Arg_2$. This is a critical constraint for the model to jointly comprehend two discourse arguments. The context c is composed of two discourse arguments Arg_1 and Arg_2, with the insertion of an explicit discourse connective $Conn_e$ or an implicit $Conn_i$ (to be inferred by the model). The question q is composed of a prompt p and a span $s_1 \in Arg_1$.

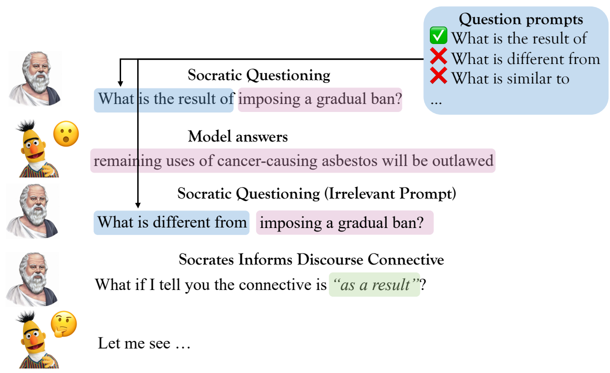


Figure 2: DISQ asks questions with all possible question prompts. When the prompt is consistent with discourse relation, a desired span should be extracted. If the prompt is irrelevant, an understanding model will abstain. DISQ also inserts a counterfactual discourse connective to guide the model to answer again.

Questioning with implicit connective: Discursive questioning *elicits* discourse relations. Our key insight is that discourse relations are a hidden variable that facilitate discursive questioning. In our running example, the model needs to understand discourse relation R as *contingency* to perform successful QA. DISQ will also ask questions with incorrect question prompts (e.g. "what is different from" question in Figure 2); an understanding model must abstain from answering these illogical questions.

Questioning with explicit connective: A realized discourse connective explicates discourse relation. We now insert a plausible discourse connective (e.g. $Conn_e$ *"as a result"* as $Conn_e$ in Figure 2) and conduct the same questioning again. Similar to how humans read, the explicit marker then assists the reader in comprehending the discourse. So if a model understands the connective and incorporates it into the comprehension of the discourse, it should perform better QA.

Question Generation: Questions are generated automatically by composing a question prompt and a span in the discourse. (1) To create a battery of question prompts, we refer to the sense taxonomy in PDTB 2.0 (Prasad et al., 2008) and produce the following prompts $\mathcal{P}_R$ for each discourse relation R in Table 2. (2) To extract proper spans, we follow previous work (Pan et al., 2020) to use a trained semantic role labeler (SRL) to find self-contained spans.

Question prompt $\mathcal{P}_R$ set	Discourse relation R
Why What is the result of What is the reason of	Contingency
What is different from What is opposite to	Comparison
What is similar to What is an example of	Expansion
What happens after What happens before	Temporal

Table 2: Question prompts and their discourse relation.

2.2 Output: DISQ's Matrices (DISQM)

DISQ's output is an array of matrices $\mathcal{M} = \{M^1, M^2, ..., M^{N-1}\}$. We name $\mathcal{M}$ as DISQ's Matrices (DISQM). Given a sequence of discourse arguments (sentences) of length $N \geq 2$, we perform DISQ in a sliding window style. M^i indicates the output for ith and $(i+1)$th sentence.

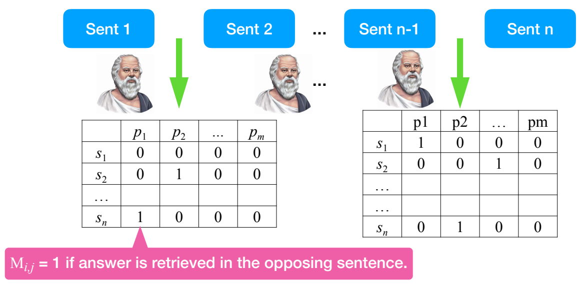


Figure 3: DISQ's Matrices (DISQM) are produced by performing DISQ in a sliding window style.

As for each M, since we do not require the ground truth Q&A pairs, we ask all possible questions. All spans s from both Arg_1 and Arg_2 are combined with all question prompts, resulting a total of $|s| \times |p|$ questions to ask. As shown in Figure 3, when asked a question composed by s_i and p_j, if an answer is retrieved from opposing argument (sentence), we assign $M_{i,j} = 1$ and $M_{i,j} = 0$ otherwise. It is simplistic because we do not discriminate between correct and wrong answers due to the lack of ground truth Q&A. But $s_1 \in Arg_1$ and $s_2 \in Arg_2$ are extracted as a cohesive tie and their relation r is characterized by question prompt p. We contribute DISQM as a new interpretable representation for discourse.

3 Task 1: Discourse Relation

We measure NLMs' understanding by how well it performs in DISQ. Our formalization is distinct from the traditional setting where accuracy for classification is the primary focus.

3.1 Formalization

Evidence Extraction: We consider (s_1, s_2, p) triplet as evidence for understanding discourse relation, and study if NLMs extract proper evidence given discourse relation R.

DISQM Value: Since we do not require (and have) the annotation for evidence triplets, we model the association between discourse relation R and question prompt p as a macro-level evaluation:

$$V(R,p) = \frac{\sum^{|M^R|} \left(\sum_{j \in \mathcal{P}_R} \sum_{i=1}^{|s|} M^R_{i,j} \right)}{|M^R|} \tag{2}$$

$V(R,p)$ is the expectation for the number of evidence triplets being retrieved using question prompt p under discourse relation R. Specifically, we perform DISQ on a corpus $C = (Arg_1, Arg_2, R, Conn)_L$ with the annotation for discourse relation and connective. $V(R,p)$ concerns all $|M^R|$ number of DISQM matrices M^R with discourse relation R. Within each M^R, we only consider columns j that correspond to R's question prompts $\mathcal{P}_R$. Finally, we consider all span s being asked equally.

Assertion 1: *$V(R,p)$ must be higher than $V(R,p')$ where $p \in \mathcal{P}_R$ and $p' \notin \mathcal{P}_R$ if a model understands discourse relation.*

Models must distinguish correct prompt p against incorrect p' under discourse relation R, which will be reflected by different DISQM values. For a random model, $V(R,p) = V(R,p')$.

3.2 Implementation Details

Dataset: We study PDTB 2.0 dataset ((Prasad et al., 2008)) because they have annotated both discourse relation and connective. We focus on implicit discourse instances because they miss discourse connectives and require non-trivial reasoning over two arguments. We perform DISQ over 2 ~20 sections in PDTB (the training split for traditional setting), including 12,362 discourse instances.

NLMs: We primarily study BERT's family, following a recent investigation about models' reasoning capacity (Choudhury et al., 2022). We experimented BERT (Devlin et al., 2019) and RoBERTa model (Liu et al., 2019) of tiny, base, and large sizes. To enable question answering, we choose BERT and RoBERTa models fine-tuned on SQuAD 2.0 dataset, which are also de facto choices for QA research. DISQ is very generic, practitioners can explore other NLMs fine-tuned on other tasks.

Evaluation Measure: We primarily study $V(R,p)$ as a proxy for understanding discourse relations. We also present a normalized $\hat{V}(R,p) = \frac{V(R,p)}{AVG(V(R,p)), R \in \mathcal{R}}$ for proper comparison among prompts. This is because we observe some prompts have a higher prior to having an answer.

3.3 Evaluation

Our evaluation is focused on the general performance on DISQ (RQ1), the role of discourse connective (RQ2), and interpretability (RQ3):

How do NLMs generally perform on DISQ? (RQ1) We first do not insert discourse connective and expect the model can understand the discourse relation. We interpret the result in Table 3 from two angles: **(1) Question(Column)-wise comparison:** There are 9 (R,p) cells we expect the highest $V(R,p)$ value in one column (bolded), for example, $V(R,p) = 0.144$ for "Comparison" question in "Different" column. We find that 7 out of the 9 **desired cells** have achieved the highest value in their columns. Interestingly, we find Expansion relation does not achieve the desired score. Our conjecture is that Expansion relation intrinsically lacks the salient semantic like *contrast* or *cause/result* in other discourse relations. **(2) Relation(Row)-wise comparison:** Normalized score $\hat{V}(R,p)$ enables relation(row)-wise comparison. We again observe the same 7 out of 9 cells achieving the highest

	Different	Opposite	Why	Result	Reason	Similar	Example	After	Before
Comparison	$\mathbf{0.144}_{\mathbf{(1.582)}}$	$\mathbf{0.022}_{\mathbf{(1.402)}}$	$0.598_{(0.884)}$	$0.642_{(0.930)}$	$0.705_{(0.880)}$	$0.154_{(0.886)}$	$0.34_{(0.814)}$	$0.453_{(0.946)}$	$0.063_{(1.042)}$
Contingency	$0.081_{(0.888)}$	$0.015_{(0.962)}$	$\mathbf{0.814}_{\mathbf{(1.204)}}$	$\mathbf{0.764}_{\mathbf{(1.108)}}$	$\mathbf{0.962}_{\mathbf{(1.200)}}$	$0.187_{(1.077)}$	$0.465_{(1.113)}$	$0.441_{(0.920)}$	$0.049_{(0.799)}$
Expansion	$0.075_{(0.820)}$	$0.015_{(0.959)}$	$0.711_{(1.051)}$	$0.633_{(0.918)}$	$0.799_{(0.997)}$	$\mathbf{0.184}_{\mathbf{(1.060)}}$	$\mathbf{0.462}_{\mathbf{(1.105)}}$	$0.380_{(0.794)}$	$0.045_{(0.748)}$
Temporal	$0.065_{(0.710)}$	$0.011_{(0.677)}$	$0.582_{(0.861)}$	$0.720_{(1.044)}$	$0.740_{(0.923)}$	$0.170_{(0.977)}$	$0.405_{(0.968)}$	$\mathbf{0.642}_{\mathbf{(1.341)}}$	$\mathbf{0.086}_{\mathbf{(1.411)}}$

Table 3: DISQ with implicit connective for BERT$_{\text{Large}}$: $V(R,p)$ is compared column-wise and $\hat{V}(R,p)$ (inside parentheses) is compared row-wise. Numbers are **bolded** if desired to be highest in its row/column and in green if achieved. 7 out of 9 cells achieve the highest value, marking a strong association between R and p (**RQ1**).

	Different	Opposite	Why	Result	Reason	Similar	Example	After	Before
Comparison	$\mathbf{2.232}_{+0.650}$	$\mathbf{2.379}_{+0.977}$	$0.666_{-0.218}$	$0.792_{-0.138}$	$0.65_{-0.230}$	$0.761_{-0.125}$	$0.652_{-0.162}$	$0.821_{-0.125}$	$0.918_{-0.124}$
Contingency	$0.552_{-0.336}$	$0.544_{-0.418}$	$\mathbf{1.433}_{+0.229}$	$\mathbf{1.331}_{+0.223}$	$\mathbf{1.627}_{+0.427}$	$1.047_{-0.030}$	$1.134_{+0.021}$	$0.853_{-0.067}$	$0.903_{+0.104}$
Expansion	$0.623_{-0.197}$	$0.660_{-0.299}$	$1.077_{+0.026}$	$0.868_{-0.050}$	$0.943_{-0.054}$	$\mathbf{1.148}_{+0.088}$	$\mathbf{1.365}_{+0.260}$	$0.664_{-0.130}$	$0.713_{-0.035}$
Temporal	$0.593_{-0.117}$	$0.417_{-0.260}$	$0.824_{-0.037}$	$1.009_{-0.035}$	$0.780_{-0.143}$	$1.043_{+0.066}$	$0.850_{-0.118}$	$\mathbf{1.662}_{+0.321}$	$\mathbf{1.465}_{+0.054}$

Table 4: DISQ replicated with explicit connectives: We report $\hat{V}(R,p)$ and Δ values compare with Table 3 (e.g. $2.232 - 1.582 = +0.650$). The performance is boosted. All 9 desired cells receive a $+\Delta$ value while the most of the undesired cells receive a $-\Delta$ value, demonstrating a strong understanding of discourse connective (**RQ2**).

score(s) in the row. For example, the Comparison relation is very responsive to "different" and "opposite" prompts ($\hat{V}(R,p)$ is 1.582 and 1.402).

Both comparisons show remarkable results for NLMs to extract evidence in consistency with discourse relation without the hint from connective. We focus on BERT$_{\text{Large}}$ model here and present other models' performance in Appendix E. All models show an association between R and p but larger models tend to perform better, which is in line with recent findings in (Choudhury et al., 2022).

Can discourse connective improve NLMs' performance? (RQ2) We then explore the effect of the counterfactual $conn_e$ which explicates the hidden variable of discourse relation. In Table 4, the normalized DISQM values $\hat{V}(R,p)$ and Δ values are presented. We find the insertion of explicit connective boosts the performance of the questioning. Now 9 out of 9 desired cells achieve the highest score in both column and row-wise comparisons (Expansion relation included). Moreover, all **desired cells** receive a $+\Delta$ value. The rest of the cells mostly receive a $-\Delta$ impact. It is remarkable for NLMs to interpret discourse by conditioning on the inserted $conn_e$ to seek more correct evidence and eliminate incorrect evidence, which we believe is similar to human-like understanding.

Case study: Is DISQ's output interpretable? (RQ3) Table 5 showcases DISQ's output on our running example. BERT$_{\text{Large}}$ model retrieves the desired answer given *"Why"* and *"What is the result of"* questions which are in line with *Contingency* relation. We also find the Q&A pairs very readable to human and contributes to discourse relation. It is worth noting that irrelevant questions like *"what is different from"* are also asked, which means the model is able to abstain from answering these questions. This example also exhibits symmetry. It is a desired structure (Topology 1) as detailed later in Section 6. *"what is the result of"* prompt extracts am answer from Arg_2, and the *"why"* prompt extracts an answer from Arg_1. As these two prompts have opposing meanings, their answers can reinforce each other symmetrically. This is similar to how people read contexts, in a bidirectional manner. We analyze additional case studies in other configurations in Appendix C.

Discourse sense: Contingency.Cause.Result. Conn: *as a result*
Arg1: In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. **Arg2:** By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.
Question: What is the result of imposing a gradual ban? **Answer:** almost all remaining uses of cancer-causing asbestos will be outlawed. **Confidence:** 0.40
Question: Why will almost all remaining uses of cancer - causing asbestos outlawed? **Answer:** the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. **Confidence:** 0.06

Table 5: Case study for BERT$_{\text{Large}}$ model with implicit connective: The model retrieves evidence only with two correct question prompts and abstains from 50+ irrelevant questions (**RQ3**).

4 Task 2: Discourse Coherence

We have studied Q&A pairs discovered by DISQ as reasoning evidence for discourse relations. We now explore such Q&A pairs as cohesive devices. We contribute DISQM values as a new reference-free measure for text coherence.

4.1 Formalization

Coherence Modeling: Given a sequence of sentences (discourse arguments) $T = \{t_1, t_2, ..., t_n\}$, a model needs to predict a coherence score V.

DISQM Values: Coherence is achieved by linking multiple spans in sentences through semantic relations (Halliday, 1976). We extend Halliday's predefined cohesion types to generic cohesion discovered by DISQ. Formally, given T, DISQ is performed on each pair of sentences, (t_1, t_2), (t_1, t_2), ... (t_{n-1}, t_n), resulting an array of DISQM matrices $\mathcal{M} = \{M^1, M^2, ..., M^{n-1}\}$. We define following DISQM aggregate values:

- $V_{sum}(\mathcal{M}) = (\sum_{k=1}^{n} \sum_{j=1}^{|P|} \sum_{i=1}^{|s|} M_{i,j}^k)/n$, (Sum)
- $V_{den}(\mathcal{M}) = (\sum_{k=1}^{n} \frac{\sum_{j=1}^{|P|} \sum_{i=1}^{|s|} M_{i,j}^k}{|P| \times |s|})/n$, (Density)
- $V_p(\mathcal{M}) = (\sum_{k=1}^{n} \sum_{j=1}^{|P|} \lceil \sum_{i=1}^{|s|} M_{i,j}^k \rceil^1)/n$, (Prompts)
- $V_s(\mathcal{M}) = (\sum_{k=1}^{n} \sum_{i=1}^{|s|} \lceil \sum_{j=1}^{|P|} M_{i,j}^k \rceil^1)/n$, (Spans)

These values are aggregations of $\mathcal{M}$ because we believe 1s in M indicate Q&A pairs which encode local cohesion, and their aggregation leads to global coherence over the discourse. The values are divided into two groups: (1) Quantity-driven: $V_{sum}(\mathcal{M})$ and $V_{den}(\mathcal{M})$ measure the average sum of the matrix M and the density of matrix M respectively. The intuition is that when more QA pairs are extracted (1s in M), more cohesive devices contribute to global coherence. (2) Diversity-driven: $V_p(\mathcal{M})$ and $V_s(\mathcal{M})$ measure the number of active question ***p***rompts and active ***s***pans in M respectively. When writers compose a context, they may use multiple discourse senses or use several cohesive devices to stress the coherence. $\lceil M \rceil^1 = clip(M, 1)$ denotes a function to clip a matrix to a max value of 1.

Assertion 2: *On average $V(T)$ should be higher than $V(T')$ when T is more coherent than T' if a model understands discourse coherence.* [2]

[2]We use T and $\mathcal{M}$ interchangeably. This assertion might have exceptions where short sentences can also be coherent but they have fewer cohesive devices.

A coherent discourse is better than random sentences because more cohesive devices link the text together. An idealist model must be able to identify them which are reflected in DISQM values.

We contribute DISQM values as a new measure for text coherence. It is simple, non-parametric, and reference-free. It is possible to exploit the topological patterns in DISQM like the Entity-grid method (Barzilay and Lapata, 2008), we now perform a qualitative study and leave it for future work.

4.2 Implementation Details

Dataset: We choose SummEval dataset (Fabbri et al., 2021) because it is a new resource providing human annotation on text coherence. They provide coherence annotation for 17 systems' output on 100 summarization instances. Notably, Fabbri et al. (2021) find that coherence is the most problematic aspect of automatic summarization evaluation (least correlated with human judgement).

4.3 Evaluation

	Sum (V_{sum})	Density (V_{den})	Spans (V_s)	Prompts (V_p)
$\text{BERT}_{\text{Tiny}}$	-0.353	-0.324	-0.279	0.0
$\text{BERT}_{\text{Base}}$	**-0.441**	**-0.382**	-0.324	**-0.382**
$\text{BERT}_{\text{Large}}$	-0.118	-0.206	0.022	0.044
$\text{RoBERTa}_{\text{Tiny}}$	-0.074	-0.088	-0.015	0.044
$\text{RoBERTa}_{\text{Base}}$	0.176	-0.074	0.324	0.338
$\text{RoBERTa}_{\text{Large}}$	**0.647**	0.294	**0.647**	**0.632**

Table 6: System-level Kendall's Tau correlation with human judgments. Scores are **bolded** if greater than or equal to previous state-of-the-art (-0.382 and 0.397 for $-$ve and $+$ve correlations (Fabbri et al., 2021))

We use $V(\mathcal{M})$ as the coherence measures. Following Fabbri et al. (2021), we use system-level Kendall's Tau correlation to assess $V(\mathcal{M})$'s correlation with human judgements.

We perform DISQ on 17×100 summarization instances and obtain their DISQM values $V(\mathcal{M})$. We report Kendall's Tau scores in Table 6 and make two observations: (1) The $\text{RoBERTa}_{\text{Large}}$ and $\text{RoBERTa}_{\text{Base}}$ models have shown a positive correlation with human judgment on coherence. Notably, the $\text{RoBERTa}_{\text{Large}}$ model even outperforms previous state-of-the-art significantly. V_{sum} has a correlation of **0.647**, significantly higher than the previous state-of-the-art). It indicates that useful cohesive devices have been extracted by NLMs, such that even our simple aggregations correlate well. (2) However, $\text{BERT}_{\text{Base}}$ show a significant negative correlation. This is counter-intuitive, as we assume that 1s in DISQM contributes positively to coherence. The cause may be due to $\text{BERT}_{\text{Base}}$ having many incorrect answers and hallucinating responses, and a consequences of BERT's fragility compared with RoBERTa. This leads us to recommend practitioners to explore larger models and architectures that exceed a minimal threshold level of performance for DISQ analyses to make sense.

4.4 Usability Tests of DISQM

We recommend two usability tests to make DISQM trustworthy and controllable. They help practitioners decide the usability of NLMs for discourse tasks. They also serve as an explanation for interesting model behaviors that we have discovered.
Test 1: Sentence ordering is an automatic usability test. Practitioners should choose models with high accuracy for this task. Itself is a classic experimental setting for coherence modeling (Lin et al., 2011). Its advantage is that it can be performed in automatically synthesized contexts. The assumption is that randomly perturbed sentences should be less coherent than the original ones.

We showcase one study on the SummEval dataset's human-written summaries. It comprises 1,000 summaries, which are all assumed to be coherent. Following the setup in (Lin et al., 2011), we generate 20 perturbations for each instance (shorter summaries may have fewer than 20 perturbations). We also follow the setting in (Lin et al., 2011) to perform a binary prediction task between original context T and perturbed context T'. We consider a prediction is correct when $V(T) > V(T')$ and incorrect otherwise.

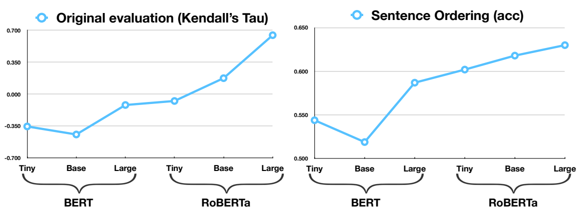


Figure 4: Usability Test 1: Models' performance on coherence modeling (left) is in a similar trend to the sentence ordering task (right).

We compare the performance of the original evaluation (measured by Kendall's Tau) and the sentence ordering task (measured by accuracy score) using V_{sum} value. We find they are in a very similar trend: $\text{BERT}_{\text{Base}}$ is the lowest and the $\text{RoBERTa}_{\text{Large}}$ the highest. RoBERTa models perform better in both tasks. Notably, the $\text{BERT}_{\text{Base}}$ model only scores 0.519 accuracy which is nearly random, meaning it cannot distinguish coherent discourse against random sentences. It explains its weak performance in the original evaluation for coherence modeling.
Test 2: Correctness of answers requires a moderate amount of human input to determine the correctness of Q&A pairs produced by DISQ. The more correct Q&A pairs, the more reliable a model is. The assumption is that only when Q&A pairs are correct, do they make a positive contribution to coherence.

Since we do not have the ground truth data for the Q&A pairs in SummEval, we manually conduct a proof-of-concept study. The first author classified the Q&A pairs into three categories: (1) **C**orrect (C): The two spans (s_1 and s_2) in question and answer satisfy the relation of the question prompt p; (2) **I**ncorrect (I): The spans are either unrelated, or do not satisfy the relation of the question prompt p; (3) **N**on-contextual (N): Two spans (s_1 and s_2) satisfy the relation of question prompt p out of context, but not in correct context. Similar definition is also adopted in (Lei et al., 2021). We randomly sample 50 summaries and study DISQ's output by $\text{BERT}_{\text{Base}}$ and $\text{RoBERTa}_{\text{Large}}$ models, which are the most negative and positive correlated models measured by Kendall's tau correlation (+0.647 and -0.441) [3]:

	C	I	N
$\text{BERT}_{\text{Base}}$	72 (24.1%)	217 (71.6%)	13 (4.3%)
$\text{RoBERTa}_{\text{Large}}$	49 (52.1%)	43 (45.7%)	2 (2.1%)

Table 7: Classification of Q&A pairs in pilot study: RoBERTa has a higher ratio of correct Q&A (52.1%).

As shown in Table 7, $\text{RoBERTa}_{\text{Large}}$ model has a much higher portion of correct answers compared to $\text{BERT}_{\text{Base}}$ model. It offers initial evidence that only correct (C) Q&A pairs are contributing to coherence and it endorses the usability test. As for the $\text{BERT}_{\text{Base}}$ model, we observe that it produces many wrong (W) and noncontextual (N) Q&A pairs. So the negative Kendall's Tau correlation might be explained in this way: incoherent context lacks an obvious or salient discourse relation so many sense seems possible. In this case, a "confused" model like $\text{BERT}_{\text{Base}}$ is likely to hallucinate and respond to many possible question prompts (We articulate our classifications with examples in Appendix D).

[3]Pilot study results are uploaded as supplementary data.

5 Related Work

QA for NLP Tasks: Even though question answering (QA) has been explored as an interface for many NLP tasks, DISQ's focus is using QA as an unsupervised approach for model interpretation. Existing works primarily explored annotating golden data and training supervised models. Notable efforts include QASRL (FitzGerald et al., 2018), QANorm (Klein et al., 2020), QADiscourse (Pyatkin et al., 2020), QASem (Klein et al., 2022), DCQA (Ko et al., 2022a), and QA for reference/ellipsis resolution (Hou, 2020; Aralikatte et al., 2021). We draw inspiration from the self-talk paradigm (Shwartz et al., 2020) that generates clarifying questions and queries NLMs for additional evidence. The key distinction is that Shwartz et al. (2020)'s answers are retrieved outside the given context, while our answer comes from the context.

Interpretation Methods in NLP: DISQ provides an unsupervised alternative to popular interpretation methods: (1) **Probing paradigm** takes out the representation of NLMs and train a model to predict whether one linguistic property is captured by the representation (Tenney et al., 2019; Wallace et al., 2019; Li et al., 2021). Despite being simple, it requires labeled data for supervision. (2) As summarized by Belinkov et al. (2020), **behavior analysis and post-hoc interpretation** produce fine-grained interpretation of model's output. The common practice is to perturb the text to reveal the decision boundary or unwanted bias of the model (Feng et al., 2018; Ribeiro et al., 2016; Poliak et al., 2018; Rudinger et al., 2018). But the creation of the perturbation usually requires human input.

Discourse Modeling: DISQ creates new possibilities for several discourse tasks: **(1) Discourse relation:** NLMs are used as a backbone for customized neural networks to predict discourse relation (Liu et al., 2016; Dai and Huang, 2018; Liu et al., 2020). Even though the performance shows improvement over prior feature-based methods (Pitler et al., 2009; Rutherford and Xue, 2014), these methods lack interpretability. One recent exception (Jiang et al., 2021) considers generation as an auxiliary task to prediction. The generated text offers some interpretability but it is not their focus. We hope future works to be evaluated and optimized by DISQ. **(2) Discourse coherence:** Similarly, neural methods (Mohiuddin et al., 2018; Jwalapuram et al., 2022) perform better than feature-based methods (Barzilay and Lapata, 2008). To the best of our knowledge, there is no existing work interpreting the inner mechanism of NLMs for coherence. We hope our formalization of Q&A pairs as cohesive devices will seed more interpretable models. **(3) Discourse structure:** Our DISQM matrices are linear, not hierarchical. We can learn from recent advances using NLMs to predict hierarchical structures (Huber and Carenini, 2022; Ko et al., 2022b; Xiao et al., 2021) .

6 Conclusion and Discussions

Due to the lack of annotated data, little progress has been made towards interpreting how NLMs understand discourse. We present the first study by enabling models to self-interrogate with a Discursive Socratic Questioning (DISQ) procedure. By analyzing DISQ's output matrices (DISQM), we find NLMs show remarkable evidence in understanding both discourse relations and coherence by identifying cohesive spans in text and realizing their relations through Socratic questioning. We urge researchers to test their NLMs with our DISQ usability tests as an additional layer of validation.

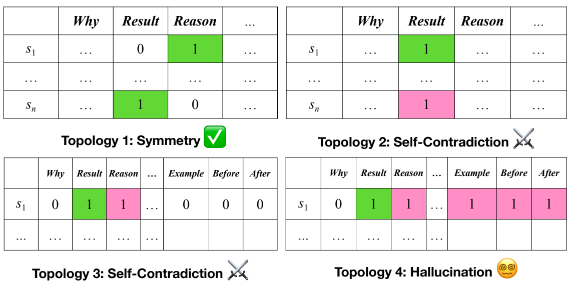


Figure 5: Symmetry, self-contradiction and hallucinations in DISQM. Green cells indicates correct answers. Pink cells indicates incorrect or noncontextual answers.

As Socratic questioning ends, students realize what they *know and do not know*. Topology 1 (Figure 5, upper left) is a DISQM result: when $s_1 \in Arg_1$ and $s_2 \in Arg_2$ are in a *reason–result* relationship with each other, a symmetric structure is established, similar to how humans read. In contrast, Topology 2 indicates self-contradiction: here, both $s_1 \in Arg_1$ and $s_2 \in Arg_2$ are considered as the result for each other, which is illogical. Finally, Topology 4 shows a model that hallucinates and responds positively to many questions, which happens when the model finds only weak relatedness. In future work, these patterns may serve as signals for self-supervision to insert logic into discursive NLMs for attaining better reliability.

Ethical Considerations and Limitations

When performing DISQ, we note that output answers may be offensive in certain contexts, because practically all spans in the context can be (incorrectly) extracted as an answer. This is a common concern for all QA models to overcome, not specific to DISQ. But according to our pilot study, we have not found any cases of such offensive Q&A pairs.

DISQ also has particular limitations. (1) We only use the behavior of the model given a set of questions as a proxy for understanding. It is not a causal analysis. We can causally study the role of individual neuron or subnetwork for discourse function in the future, similar to a recent study about individual neuron's role for factual knowledge (Meng et al., 2022). (2) Our method is unsupervised and does not require ground-truth QA pairs. It is meaningful to create such a dataset with ground truth QA pairs annotated for discourse understanding and benchmark how models perform reasoning on it. (3) We have only studied standard English corpora. It is meaningful to apply DISQ to NLMs' understanding of discourse on other English corpora with language variations and to corpora in other languages.

References

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.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 1–5.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. 2022. Machine reading, fast and slow: When do models" understand" language? *COLING*.

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

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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. *arXiv preprint arXiv:1805.05377.*

MAK Halliday. 1976. Cohesion in english. *Longman.*

Michael Alexander Kirkwood Halliday, Christian MIM Matthiessen, Michael Halliday, and Christian Matthiessen. 2014. *An introduction to functional grammar*. Routledge.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438.

Patrick Huber and Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *NAACL.*

Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. 2021. Not just classification: Recognizing implicit discourse relation on joint modeling of classification and generation. In *Proceedings of* *the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431.

Prathyusha Jwalapuram, Shafiq Joty, and Xiang Lin. 2022. Rethinking self-supervision objectives for generalizable coherence modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6044–6059.

Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022. Qasem parsing: Text-to-text modeling of qa-based semantics. *arXiv preprint arXiv:2205.11413*.

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, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei-Jen Ko, Cutter Dalton, Mark Simmons, Eliza Fisher, Greg Durrett, and Junyi Jessy Li. 2022a. Discourse comprehension: A question answering framework to represent sentence connections. *EMNLP*.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2022b. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. *arXiv preprint arXiv:2210.05905*.

Wenqiang Lei, Yisong Miao, Runpeng Xie, Bonnie Webber, Meichun Liu, Tat-Seng Chua, and Nancy F Chen. 2021. Have we solved the hard problem? it's not easy! contextual lexical contrast as a means to probe neural coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13208–13216.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically evaluating text coherence using discourse relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 997–1006.

Xin Liu, Jiefu Ou, Yangqiu Song, and Xin Jiang. 2020. On the importance of word and sentence representation learning in implicit discourse relation classification.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*.

Tasnim Mohiuddin, Shafiq Joty, and Dat Tien Nguyen. 2018. Coherence modeling of asynchronous conversations: A neural entity grid approach. *arXiv preprint arXiv:1805.02275*.

Christine A Padesky. 1993. Socratic questioning: Changing minds or guiding discovery. In *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, volume 24.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1463–1475.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *NAACL-HLT (2)*.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629.

Samson Tan, Shafiq Joty, Kathy Baxter, Araz Taeihagh, Gregory A Bennett, and Min-Yen Kan. 2021. Reliability testing for natural language processing systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4153–4169.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in embeddings. 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 5307–5315.

Wen Xiao, Patrick Huber, and Giuseppe Carenini. 2021. Predicting discourse trees from transformer-based neural summarizers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4139–4152.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*.

A DISQ's Possible Extension to Other NLP Tasks

In this paper, we propose a self-interrogation procedure (DISQ) to interpret models' decision processes for discourse understanding. We describe a conceptual extension of DISQ that can be applied to other NLP tasks for unsupervised interpretation of models' decision process. This extension follows our two-step design:

Step 1: Socrates Asks: (1) Span identification: We first identify key spans to compose the questions. The linking of the spans may have different functions in different tasks. In discourse, we have explored the spans' linkage as cohesive devices. In Natural Language Inference (NLI), for example, two spans may compose an entailment or contradiction relation (Camburu et al., 2018). **(2) Question generation** We then generate questions with predefined question prompts customized for each task. In the NLI task, such question prompt can be *"What results in"* and *"What contradicts"*.

Step 2: Model Answers: We interrogate the model with the battery of questions automatically generated in Step 1. In line with our measure for discourse, we use the model's behavior in the questioning as a proxy for its understanding of the task. For example, in an *"entailment"* NLI instance, the model needs to answer consistently with the *"entailment"* relation. That is to say, it must extract a correct span in hypothesis with *"What results in"* prompt and abstain from *"What contradicts"* prompt.

We now briefly discuss how DISQ's extension can be applied to natural language inference (relation classification for (two sentences), sentiment analysis (single sentence classification), and text summarization (text generation):

- **Natural language inference (NLI):** As the example in Figure 6, there are two highlighted spans that signals the contradiction relation between the premise and hypothesis. We believe the model must answer correctly to *"What contradicts with"* question and abstain from other questions to have a good understanding. This example is excerpted from the e-SNLI (Camburu et al., 2018) corpus for explainable NLI. This corpus requires great effort for annotation, our DISQ can alleviate it by automatically identifying spans and generating thoughtful questions.

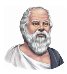
Tasks	Example	 Socrates Asks (Step 1)	 Model Answers (Step 2)
Natural Language Inference	**Premise:** An adult dressed in black holds a stick. **Hypothesis:** An adult is walking away, empty-handed. **Label:** contradiction	Q: What contradicts with holding a stick?	A: empty-handed.
Sentiment Analysis	**Input:** visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride from innocence to experience without even a hint of that typical kiddie-flick sentimentality. **Label:** Positive	Q: What is happy in the context?	A: visually imaginative, thematically instructive and thoroughly delightful
Summarization	**Reference:** Paul Merson, the Sky Sports pundit, criticized Andros Townsend ***(s1)*** last week after his call-up to the england squad. Merson admitted it was a mistake after Townsend scored, bringing the match against Italy to a tie ***(s2)*** on Tuesday. Merson is a former Arsenal player himself. **Generation:** paul merson criticised andros townsend ***(s3)***'s call-up to the england squad . townsend hit back at merson after scoring for england against italy ***(s4)*** . the tottenham midfielder was brought on in the 83rd minute against burnley .	Q1: What happens after Paul Merson, the Sky Sports pundit, criticized Andros Townsend ***(s1)***? Q2: What happens before Townsend scored, bringing the match against Italy to a tie ***(s2)***?	A1: townsend hit back at merson after scoring for england against italy ***(s4)*** A2: paul merson criticised andros townsend ***(s3)***

Figure 6: DISQ can be extended to perform unsupervised model interpretation on other NLP tasks. **Step 1:** We automatically generate Socratic-style questions with pre-defined prompts (in blue) and Spans in context (in purple). **Step 2:** Models are interrogated with these questions and we measure how well models perform in the questioning.

- **Sentiment analysis:** We believe the highlight span in Figure 6 is the evidence for a positive sentiment. A model needs to identify it with *"What is happy"* question prompt. Sentiment analysis, as a single sentence classification, may only require one span as the evidence, which is different from the reasoning over multiple spans in discourse and NLI. Therefore there might be no spans used in the questions.

- **Text summarization:** Recent papers have initially studied using QA as a new measure for summarization evaluation. They generate a question from the reference summary and query the generated summary. However, they have not explored the role of discourse in their method. We briefly discuss how DISQ can incorporate discourse semantics into using QA for summarization evaluation. As shown in the reference summary in Figure 6, the two spans s_1 and s_2 link the discourse together with a salient *Temporal* relation. We believe such a relation is the key to making the summary coherent and should be reserved in the generated summary. We show a good generated summary where s_3 and s_4 also express such *Temporal* relation. We generate **Question 1** with *"What happens after"* prompt and s_1, expecting the answer s_4 from the generated summary. In the meantime, **Question 2** combines *"What happens before"* prompt and s_2 and we expect its answer s_3 from the generated summary. If the model can answer correctly for both questions, we believe the *Temporal* relation is realized in the generated summary. Interestingly, the reference and generation exhibit a symmetric property ($s_1 - s_4$ and $s_2 - s_3$). It is in the same spirit as we desire a good discourse understanding.

B DISQ Generalizes Halliday's Cohesion Theory

We contribute DISQ as a computational tool to discover new cohesive ties. We recommend linguists apply DISQ on their corpora and examine the output. Halliday (1976)'s cohesion ties are well-defined but constrained. DISQ loosens these constraints by considering arbitrary semantic relation between arbitrary spans, conditioning on discourse relation and discourse coherence. NLMs are powerful tools by modeling the co-occurrence between words and sentences on billion texts, they have the

Cohesive Tie	Example	Semantic Relation	Common Spans in the Tie
Conjunction	Someone comes along with a great idea for an expedition, for example, I did a book called Sand Rivers, just before the Indian books, and it was a safari into a very remote part of Africa.	/	/
Reference – **Exophoric reference**	Kate I must say this fish is cooked beautifully.	Identical	[Nominal, adverbial group] $\sim$ [Environment]
Reference – **Endophoric reference**	There was once a velveteen rabbit. He was fat and bunchy ...	Identical / similar / exclusive	[Nominal, adverbial group] $\sim$ [Word (*he*, *it*)]
Substitution	Is he at home? I think so.	yes/no	[Clause, nominal, adverbial group] $\sim$ [Word (*so*, *do*)]
Ellipsis	Is he at home? Yes he is $\emptyset$: at home.	yes/no	[Clause, nominal, adverbial group] $\sim$ [$\emptyset$]
Lexical Cohesion	... have you ever heard of any other kinds of literature in the medieval period besides Chaucer?	Lexical relation (e.g. synonymy, hypernymy)	[word] $\sim$ [word]
DiSQ (Ours)	In July, the Environmental Protection Agency imposed a gradual ban, virtually all uses of asbestos. By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.	Arbitrary relation (e.g. causal, comparative, similar, temporal)	Arbitrary span (Word, SRL-based spans, nominal and adverbial groups, clauses)

Table 8: Comparing DiSQ with a non-exhaustive summary of (Halliday, 1976)'s cohesive ties. [·]∼[·] denotes two spans forming a cohesive tie. DiSQ covers a wider range of semantic relations and allows longer spans to be considered for cohesion. Some examples are excerpted from Ch. 9 in (Halliday et al., 2014).

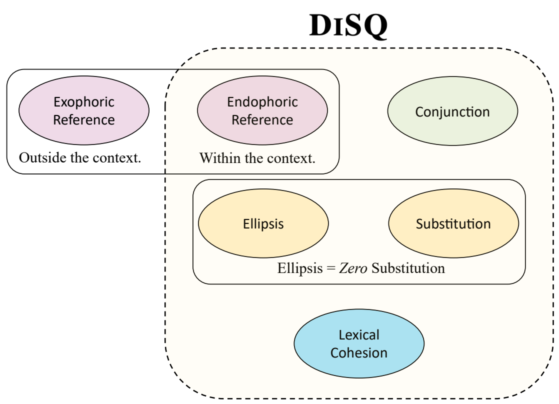


Figure 7: DiSQ is a generalized extension to Halliday (1976)'s cohesion theory. Most of the defined cohesion types can be realized by DiSQ, with the exception of exophoric reference which points outwards the text.

potential to inspire new cohesion theory.

DiSQ generalizes Halliday's theory in two aspects: **(1) Semantic relation:** DiSQ enlarges the space of semantic relation for cohesion by the unlimited choice of question prompts. We have explored causal, comparative, equivalent, and temporal semantic relation using textual (discrete) prompt in this work. In future, it is interesting to design soft (continuous) prompts by fusing different semantic relations. However, Halliday's cohesion theory only covers a very limited set of semantic relations, for example, identical and exclusive relation for reference, yes/no relation for ellipsis. The only exception is lexical cohesion. Richer lexical relations (synonymy, hypernymy, hyponymy) are the cohesive force. But it only operates on lexical items without considering longer textual units. **(2) Spans in the tie:** DiSQ is able to explore the semantic relation between arbitrary spans. Even though we only studied SRL-based spans, it is easy to adapt to other spans like lexical items, nominal groups, and clauses to realize Halliday's cohesive ties. However, Halliday's ties are much more constrained than ours. They either work between words (lexical cohesion) or between one longer span and another word (pronouns like *he* or auxiliary like *do*). With the help of DiSQ, we can explore the cohesive ties between two longer spans.

We now briefly summarize each type of Halliday's cohesion ties in Table 8 and discuss how they can be generalized by DiSQ. **(1) Conjunction**:

Halliday defines conjunctions as markers that link clauses cohesively. It is very similar to discourse connectives that link discourse arguments (some are longer sentences) together. We highlight the role of connective (conjunction) in DISQ and offer linguists a tool to test its function computationally. **(2) Reference**: Unlike conjunction that links whole clauses, reference achieves cohesion by linking elements in clauses. There are two types of references. Exophoric reference points outwards from the text and links to the environment the speakers and readers share. DISQ cannot handle such cases because we seek answers in context. Endophoric reference links elements in context. But we find the semantic relations are much more constrained to express the referential relation and the spans usually include words like personal pronouns. Longer forms of reference have been overlooked. **(3) Substitution** and **(4) ellipsis** are functionally equivalent since ellipsis can be considered as zero substitution. The cohesion is achieved through a (zero) substituted text span. Similar to reference, we find the semantic relation and spans are constrained to a small set. **(5) Lexical cohesion:** Unlike previous cohesive devices working at the grammatical level, lexical cohesion works at the lexical level by the choice of words. Even though they cover richer lexical semantics, they are constrained to work on word pairs. **(6) DISQ** is a generalized extension for Halliday's theory. It models cohesion through arbitrary semantic relations between arbitrary spans. DISQ offers a computational estimation for the effects of conjunction, and it can realize reference, substitution, ellipsis, and lexical cohesion with simple adaptations. Recently Hou (2020); Aralikatte et al. (2021) have studied to approach reference and ellipsis through QA. DISQ can extend this line and explore a wider range of cohesive devices computationally.

C Case study for Discourse Relation (Task 1)

We only present one case study in Section 3. We now analyze more cases from the PDTB dataset to examine (1) whether the output is interpretable for human; (2) whether the answers are consistent with discourse relation. Specifically, we demonstrate one more successful case with the help of counterfactual explicit connective. We also present unsuccessful cases where undesired QA pairs are extracted. Finally, we present a curious case that possibly improves the prediction of discourse sense by inserting plausible connectives. We choose the $\text{BERT}_{\text{Large}}$ model because it has achieved good overall performance in DISQ.

Discourse sense: Contingency.Cause.Result. Conn: *as a result*
Arg1: In July, the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. **Arg2:** By 1997, almost all remaining uses of cancer-causing asbestos will be outlawed.
Question: What is the result of imposing a gradual ban? **Answer:** almost all remaining uses of cancer-causing asbestos will be outlawed. **Confidence:** 0.58 (+0.18)
Question: What happens after imposing a gradual ban? **Answer:** almost all remaining uses of cancer-causing asbestos will be outlawed. **Confidence:** 0.20 (–)
Question: Why will almost all remaining uses of cancer - causing asbestos outlawed? **Answer:** the Environmental Protection Agency imposed a gradual ban on virtually all uses of asbestos. **Confidence:** 0.18 (+0.12)

Table 9: Successful cases for DISQ with connective: Desired question prompts have retrieved their answers. Their confidence scores are even increased compared to the case without connective in Table 5. One more question prompt, *"What happens after"*, has retrieved its answer.

Successful Case for DISQ with Explicit Connective Table 9 presents the same example in Table 5. The only difference is we insert the discourse connective between two arguments. We can observe that all desired questions in Table 5 can retrieve their answers. The additional question retrieving answer is *"What happens after"*. We don't count it as incorrect because the meanings of "What is the result" and "What happens after" are similar to each other and this QA pair is interpretable by us. Interestingly, we find the confidence scores are even increased as compared to Table 5. It is interesting to explore the effect of the answer's confidence in future work.

Unsuccessful Case for DISQ with Counterfactual Connective Table 10 presents a failure case for DISQ. The ground-truth discourse relation is Expansion, but we can see the question prompts are blurred together even if we have inserted the discourse connective. Both Expansion senses and Contingency senses are indicated. However, we do not attribute this failure entirely to the limitation of the BERT model's capacity. We can feel the discourse sense between Arg1 and Arg2 is indeed very ambiguous in Table 10. If this is the case, once DISQ has a blurred response, it might indicate the intrinsic ambiguity of the discourse it is processing.

Discourse sense: Expansion Conn: *in other words*
Arg1: that these events took place 35 years ago **Arg2:** It has no bearing on our work force today
Question: What is the result of taking place? **Answer:** It has no bearing on our work force **Confidence:** 0.08
Question: Why did it have no bearing on our work force? **Answer:** these events took place 35 years ago **Confidence:** 0.46
Question: What is the reason of having no bearing on our work force? **Answer:** these events took place 35 years ago **Confidence:** 0.61
Question: What is similar to having no bearing on our work force? **Answer:** these events took place 35 years ago **Confidence:** 0.20
Question: What is an example of having no bearing on our work force? **Answer:** these events took place 35 years ago **Confidence:** 0.30

Table 10: Unsuccessful cases for DISQ with connective: The questions prompts are blurred even if we insert the discourse connective. They point to both Contingency and Expansion senses.

Discourse sense: Comparison
Arg1: One claims he's pro-choice. **Arg2:** The other has opposed a woman's right to choose.
Probability of predicted discourse sense: Comparison: 0.42, Expansion: 0.49
Insert "however" as a plausible discourse connective. **# of answers:** +2
Insert "in addition" as a plausible discourse connective. **# of answers:** +0

Table 11: Curious cases for DISQ: It is possible to exploit the predictive power of DISQ to benefit the prediction task of discourse sense. We can insert plausible discourse connective and exploit the changes of DISQ's output for better sense prediction.

Curious Case of Using DISQ to help prediction Finally, we discuss a curious case of extending DISQ as an interpretation method to a predictive tool. As shown in Table 11, the prediction model is hesitating at the decision boundary for Comparison or Expansion relation. Now we insert the discourse connective for both plausible predicted senses: *"however"* for Comparison sense and *"in addition"* for Expansion sense. We observe that the model is able to generate two more answers after the insertion of *"however"* and no more answers for *"in addition"*. It is possible to formalize this intuition as an iterating process: (1) we first insert a plausible connective to perform DISQ; (2) we then leverage DISQ's output to predict discourse sense and map it back to the connective. We leave this interesting exploration for future work.

Summary of the Case Study: We perform an instance-level case study on DISQ's process on the PDTB dataset. We find those desired QA pairs are interpretable by human (performed only by the author as a case study). We also identify undesired QA pairs in discourse. The reason might be attributed to both the limitation of NLMs and the intrinsic ambiguity of the discourse senses. We conclude with a curious case of exploiting DISQ for its potential predictive power.

D Case Study for Discourse Coherence (Task 2)

We have explored using DISQ's matrices (DISQM) for coherence modeling. We observe both positive and a negative correlation with human's judgement in Section 4. We explain it by an assertion that only correct Q&A pairs discovered by DISQ make a positive contribution to coherence. We now articulate our criteria for classifying Q&A pairs and showcase real DISQM generated from the SummEval dataset.

D.1 Criteria for Classifying Q&A Pairs

Example 1:
Sent$_1$: a mother was holding the two-year-old boy and another child when the toddler slipped and fell into the pit at 3pm on saturday. **Sent$_2$:** his parents jumped in and pulled him to safety before paramedics arrived to treat the boy for a leg injury.
Ex. 1.1, Correct **Q1:** What happens after falling into the pit? **A1:** his parents jumped in and pulled him to safety
Ex. 1.2, Incorrect (Type 1) **Q2:** What happens before falling into the pit? **A2:** his parents jumped in and pulled him to safety
Ex. 1.3, Incorrect (Type 2) **Q3:** What is the reason of his parents jumping in? **A3:** 3pm on saturday.
Example 2: **Sent$_1$:** luigi costa, 71, is accused of killing his elderly neighbour terrence freebody in the dining room of his home on mugga way, red hill, canberra in july 2012. **Sent$_2$:** forensic psychiatrist professor paul mullen examined costa after the attack and believes there was evidence of the accused's state of mind declining .
Ex. 2, Non-contextual **Q1**: What is the result of killing his elderly neighbour terrence freebody? **A1:** state of mind declining

Table 12: Criteria for classifying Q&A pairs into correct, incorrect (two types), and non-contextual categories. Examples are excerpted from DISQ's output on the SummEval dataset.

DISQ links spans in discourse through the Q&A pairs extracted by questioning. Due to the limitation of NLMs, only a portion of extracted Q&A pairs is correct. We classify them with the following criteria:

- **Correct:** The two spans (s_1 and s_2) in question and answer satisfy the relation of the question prompt p. The link between s_1 and s_2 contributes to coherence and is the key to understanding discourse relation. As Ex. 1 in Table 12, the semantic relation indicated by *"what happens after"* is the key to understanding the temporal relation.

- **Incorrect:** There are two types of incorrect cases. **Type 1, Incorrect prompt:** s_1 and s_2 are related, but their relation is not consistent with the question prompt p. The two spans in Ex. 2 are indeed related, but their relation is not indicated by *"what happens before"*. **Type 2, Irrelevant spans:** s_1 and s_2 are not related. That is to say, the model retrieves a wrong answer. As in Ex. 3, the two spans are not relevant and should not be retrieved by the model.

- **Non-contextual:** Two spans (s_1 and s_2) satisfy the relation of question prompt p out of context, but not in correct context. Let's study Ex. 2 in Table 12, it is reasonable to consider "state of mind declining" as the result of the victim of a murder. But in the given discourse, "state of mind declining" actually refers to the murderer, hence the two spans do not satisfy the *"result"* relation.

D.2 DISQM from SummEval Dataset

We demonstrate DISQ's output matrices (DISQM) given instances in SummEval dataset. We cover the four topologies we discussed in Section 6, with desired symmetric properties, and undesired properties including self-contradiction and hallucination.

	Why	*Result*	*Reason*	*Different*	*Opposite*	*Similar*	*Example*	*Before*	*After*
are	0	0	1	0	0	0	0	0	0
look	0	1	0	0	0	0	0	0	0

Sent1: barcelona are six points clear at the top of la liga.
Sent2: luis enrique only took charge of the club last summer.

Q1: What is the reason of being six points clear at the top of la liga ?
A1: luis enrique only took charge of the club last summer

Q2: What is the result of taking charge of the club ?
A2: barcelona are six points clear at the top of la liga

Figure 8: DISQM and Q&A pairs for symmetric structure (**Topology 1**).

Topology 1 (Symmetry): Figure 8 shows a symmetric structure emerges in DISQM. The two spans come from the two opposing sentences, and they can extract each other as the answer with opposing prompts (*"result"-"reason"*). Even though we feel the causal semantic is not as strong as the temporal relation (characterized by *"what happens before/after"* prompts), we still recognize the model as being self-consistent and reinforce its comprehension with such a symmetric structure.

	Why	*Result*	*Reason*	*Different*	*Opposite*	*Similar*	*Example*	*Before*	*After*
well	1	0	0	0	0	0	0	0	0
race	0	0	0	0	0	0	0	0	0
finish	1	0	0	0	0	0	0	0	0

Sent1:wiggins will race in front of a sell-out crowd at london's olympic velodrome .
Sent2: the briton finished his team sky career at paris-roubaix last sunday .

Q1: Why will wiggins race in front of a sell - out crowd at london 's olympic velodrome ?
A1: the briton finished his team sky career

Q2: Why did the briton finish his team sky career at paris - roubaix last sunday ?
A2: wiggins will race in front of a sell-out crowd

Figure 9: DISQM and Q&A pairs for self-contradiction case (**Topology 2**).

Topology 2 (Self-contradiction): Self-contradiction emerges in the DISQM in Figure 9. Two spans, s_1 and s_2 are extracting each other as the answer with the same prompt *"why"*. It means the model believes s_1 and s_2 are reasons for each other. Such circular reasoning is considered self contradiction and demonstrates that the model has not fully understood the discourse.

	Why	*Result*	*Reason*	*Different*	*Opposite*	*Similar*	*Example*	*Before*	*After*
appealing	0	0	0	0	0	0	0	0	0
help	0	1	1	0	0	0	0	1	1
identify	0	1	1	0	0	0	1	1	1
robbed	1	1	0	0	0	0	1	1	1
made	0	0	0	0	0	0	0	0	0

Sent1:new zealand police are appealing to the public to help identify a man who robbed a christchurch dairy.
Sent2: he made off with the dairy 's till and about $ 1500 in cash.

Q1: Why did a man rob a christchurch dairy ?
A1: he made off with the dairy 's till and about $ 1500 in cash

Q2: What is the result of robbing a christchurch dairy ?
A2: $ 1500 in cash

Q3: What is an example of robbing a christchurch dairy ?
A3: the dairy 's till

Q4: What happens after robbing a christchurch dairy ?
A4: he made off with the dairy 's till and about $ 1500 in cash

Q5: What happens before robbing a christchurch dairy ?
A5: the dairy 's till and about $ 1500 in cash

Figure 10: DISQM and Q&A pairs for self-contradiction case (**Topology 2**) and hallucination (**Topology 4**).

Topology 3 (Self-contradiction) and Topology 4 (Hallucination): Let's now focus on the fourth

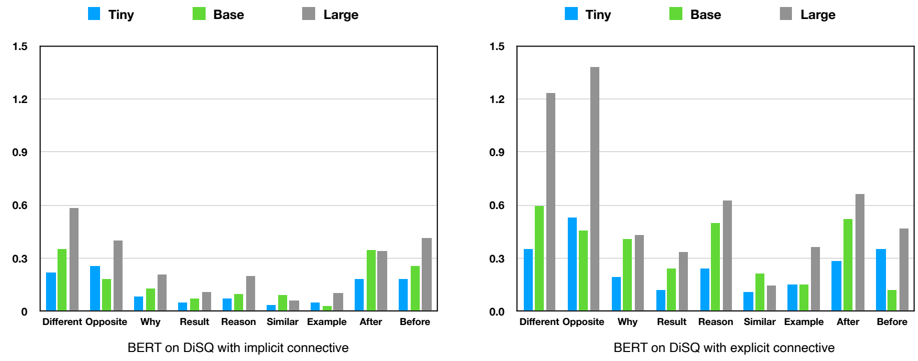


Figure 11: Performance of BERT models on DISQ: We present the score of $\hat{V}(R, p) - 1$. A score > 0 is considered sensitive between R and p. There is a steady trend that sensitivity increases with model size increases (Tiny $\rightarrow$ Base $\rightarrow$ Large). Insertion of explicit connectivity (right) boosts the association for all models without connective (left).

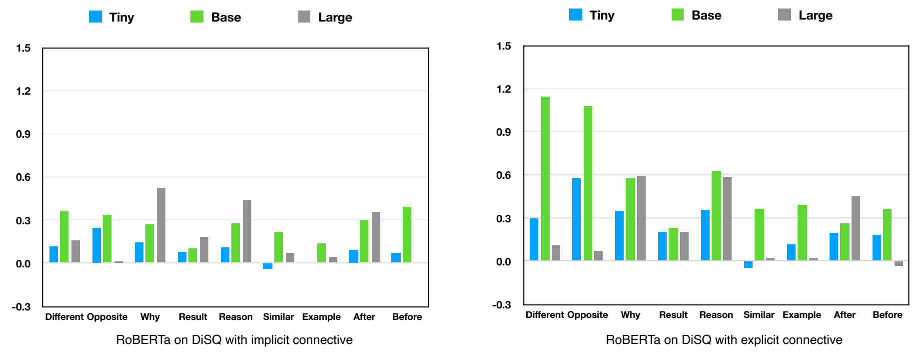


Figure 12: Performance of RoBERTa on DISQ: We present the score of $\hat{V}(R, p) - 1$. A score > 0 is considered sensitive between R and p. A steady trend among Tiny, Base, Large model sizes can be observed for DISQ without connective (left), but the tread is not clear in DISQ with connective (right). The Y-axis is in the same scale as Figure 11.

row in the DISQM in Figure 10. We only showcase DISQ's output given the span of "robbing a christchurch dairy". (1)Self-contradiction: We first find that both *"why"* and *"result"* extracts similar answers in the meantime, which is not logical. A similar case also happens in *"before"* and *"after"* question prompts, which is not logical because a fact cannot happen before and after another fact in the meantime. (2) Hallucination: Model responds to 5 out of 9 question prompts. Besides the illogical cases discussed already, the model also retrieves an incorrect Q&A pair using the *"example"* prompt. It might be explained by a conjecture that the model may not infer the discourse relation properly and decides whether many spans are related to each other.

E How do NLMs' different designs impact DISQ?

We have walked through fine-grained studies for one model in Section 3, let's now compare different models' performance on DISQ. This is an interesting question because different models can lead to different performances on discourse tasks ((Chen et al., 2019)). To facilitate inter-model comparison, we simplify the measure of $sen_n(R, p)$ by only considering $p \in \mathcal{P}_R$, for which we desire a high sensitivity (*i.e.*, those desired cells marked **as bolded**). That is to say, we approximate a column's result by only one cell (the desired cell marked **as bolded**). For example, in the column of *"What is different from"* question in Table 4, we approximate it by the cell of *Comparison* relation, which represents $\hat{V}(R, p) = 2.232$. In practice, we present $\hat{V}(R, p) - 1$, because a random baseline should also achieve a normalized $V(R, p)$ of 1.

We first present how BERT models ((Devlin et al., 2019)) of different sizes perform on DISQ. We have obtained the following findings: (1) Clearly all models demonstrate association between R and p. It is a strong result for BERT models to comprehend discourse by reasoning over spans and identifying the relations among them. The prompts for Comparison and Temporal relations are performing better, which is in line with our discovery in Section 3. (2) Figure 11 exhibits a clear correlation between performance and model sizes. The Tiny model (blue) tends to be least sensitive, the Base model (green) the medium, and the Large model (grey) the most sensitive. This trend is steady *w.r.t.* different question prompts. (3) Insertion of explicit connective also brings steady boosts to almost all models and all questions. Interesting, there is a big variance w.r.t. the boost to different questions. For example, *"What is different from"* questions are much higher than *"What is similar to"* questions. This might be due to the frequency of keywords in pre-training data or fine-tuning data ((Razeghi et al., 2022)).

We then perform the same set of experiments on RoBERTa models ((Liu et al., 2019)). We have the following findings: (1) RoBERTa models also have a good performance on DiSQ. As we can see in Figure 12, most sensitivity scores are positive values, and only a small portion of them have slightly negative values. It is a strong result for RoBERTa, because it has removed the Next Sentence Prediction (NSP) training objective which is believed to be useful for modeling over longer contexts. It means RoBERTa has implicitly constructed discourse-level understanding through other training objectives; (2) We find the trend among Tiny, Base, and Large is steady for DiSQ without connective (left), but not steady for DiSQ with connective (right). For example, the Base model achieves better $V(R, p)$ than the Large model in "Different" and "Opposite" questions in the right figure. We leave the exploration of this interesting phenomenon for future work.

F Reproducibility

F.1 Neural Language Models (NLMs)

We have applied DiSQ to examine NLMs' capacity for discourse understanding. We follow Choudhury et al. (2022) to study BERT family (Devlin et al., 2019; Liu et al., 2019).

To enable NLMs to perform QA, we choose models fine-tined on SQuAD 2.0 dataset (Rajpurkar et al., 2018). Specifically, we use a set of off-the-shelf models shared through the Hugging Face community. This is because these models are very popular in the community and many applications have been built on top of them. We hope our findings generated with DiSQ can help the users of these models diagnose the discourse capacity for these models. These models also come with MIT or CC BY 4.0 licenses. Detailed model cards can be found in the URLs.

Model	Configurations	URL
BERT$_{tiny}$	67M parameters, 6*l*, 768*d*	https://huggingface.co/deepset/tinybert-6l-768d-squad2
BERT$_{base}$	110M parameters, 12*l*, 768*d*	https://huggingface.co/deepset/bert-base-uncased-squad2
BERT$_{large}$	340M parameters, 24*l*, 1024*d*	https://huggingface.co/deepset/bert-large-uncased-whole-word-masking-squad2
RoBERTa$_{tiny}$	76M parameters, 6*l*, 768*d*	https://huggingface.co/deepset/tinyroberta-squad2
RoBERTa$_{base}$	125M parameters, 12*l*, 768*d*	https://huggingface.co/deepset/roberta-base-squad2
RoBERTa$_{large}$	355M parameters, 24*l*, 1024*d*	https://huggingface.co/deepset/roberta-large-squad2

Table 13: We examine BERT family with different configurations. Please refer to the URL for model details.

F.2 Computational Costs

DiSQ is an unsupervised interpretative measure, hence no training is required. We directly deploy the off-the-shelf NLMs and do not tune any parameters of it. As for the evaluation of the PDTB dataset (around 12k discourse instances), the computation costs around 3 hours, 6 hours, and 10 hours for tiny, base, and large models respectively on a single NVIDIA V100 GPU. As for the evaluation of the SummEval dataset (1700 summaries), the computation costs around 5 hours, 10 hours, and 17 hours for tiny, base, and large models respectively on a single NVIDIA V100 GPU.

F.3 Packages

We use AllenNLP's toolkit for semantic role labeling [4] for question generation and use spaCy model[5] to perform sentence segmentation for the summaries in the SummEval dataset.

[4] https://storage.googleapis.com/allennlp-public-models/structured-prediction-srl-bert.2020.12.15.tar.gz

[5] https://github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.4.0