QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines

Valentina Pyatkin[1] **Ayal Klein**[1] **Reut Tsarfaty**[1,2] **Ido Dagan**[1]
[1]Computer Science Department, Bar Ilan University
[2]Allen Institute for Artificial Intelligence
{valpyatkin,ayal.s.klein,ido.k.dagan,reut.tsarfaty}@gmail.com

Abstract

Discourse relations describe how two propositions relate to one another, and identifying them automatically is an integral part of natural language understanding. However, annotating discourse relations typically requires expert annotators. Recently, different semantic aspects of a sentence have been represented and crowd-sourced via question-and-answer (QA) pairs. This paper proposes a novel representation of discourse relations as QA pairs, which in turn allows us to crowd-source wide-coverage data annotated with discourse relations, via an intuitively appealing interface for composing such questions and answers. Based on our proposed representation, we collect a novel and wide-coverage QADiscourse dataset, and present baseline algorithms for predicting QADiscourse relations.

1 Introduction

Relations between propositions are commonly termed *Discourse Relations*, and their importance to the automatic understanding of the content and structure of narratives has been extensively studied (Grosz and Sidner, 1986; Asher et al., 2003; Webber et al., 2012). The automatic parsing of such relations is thus relevant to multiple areas of NLP research, from extractive tasks such as document summarization to automatic analyses of narratives and event chains (Li et al., 2016; Lee and Goldwasser, 2019).

So far, discourse annotation has been mainly conducted by experts, relying on carefully crafted linguistic schemes. Two cases in point are PDTB (Prasad et al., 2008; Webber et al., 2019) and RST (Mann and Thompson, 1988; Carlson et al., 2002). Such annotation however is slow and costly. Crowd-sourcing discourse relations, instead of using experts, can be very useful for obtaining larger-scale training data for discourse parsers.

The executions were **spurred** by lawmakers **requesting** action to **curb** rising crime rates.
What is the reason lawmakers **requested** action? to **curb** rising crime rates *What is the result* of lawmakers **requesting** action to **curb** rising crime rates? the executions were **spurred**
I **decided** to do a press conference [...], and I **did** that going into it **knowing** there would be consequences.
Despite what did I **decide** to do a press conference? **knowing** there would be consequences

Table 1: Sentences with their corresponding Question-and-Answer pairs. The bottom example shows how *implicit* relations are captured as QAs.

One plausible way for acquiring linguistically meaningful annotations from laymen is using the relatively recent QA methodology, that is, converting a set of linguistic concepts to intuitively simple Question-and-Answer (QA) pairs. Indeed, casting the semantic annotations of individual propositions as narrating a QA pair gained increasing attention in recent years, ranging from QA driven Semantic Role Labeling (QASRL) (He et al., 2015; FitzGerald et al., 2018; Roit et al., 2020) to covering all semantic relations as in QAMR (Michael et al., 2018). These representations were also shown to improve downstream tasks, for example by providing indirect supervision for recent MLMs (He et al., 2020).

In this work we address the challenge of crowd-sourcing information of higher complexity, that of discourse relations, using QA pairs. We present the QADiscourse approach to representing intra-sentential Discourse Relations as QA pairs, and we show that with an appropriate crowd-sourcing setup, complex relations between clauses can be effectively recognized by non-experts. This layman annotation could also easily be ported to other languages and domains.

Specifically, we define the QADiscourse task to be the detection of the two discourse units, and the

labeling of the relation sense between them. The two units are represented in the question body and the answer, respectively, while the question type, as expressed by its prefix, represents the discourse relation sense between them. This representation is illustrated in Table 1 and the types of questions we focus on are detailed in Table 3. This scheme allows us to ask about both *explicit* and *implicit* relations. To our knowledge, there has been no work on collecting such question types in a systematic way.

The contribution of this paper is thus manifold. (i) We propose a novel QA-based representation for discourse relations reflecting a subset of the sense taxonomy of PDTB 3.0 (Webber et al., 2019). (ii) We propose an annotation methodology to crowdsource such discourse-relations QA pairs. And, (iii) given this representation and annotation setup, we collected QADiscourse annotations for about 9000 sentences, resulting in more than 16600 QA pairs, which we will openly release. Finally, (iv) we implement a QADiscourse parser, serving as a baseline for predicting discourse questions and respective answers, capturing multiple discourse-based propositions in a sentence.

2 Background

Discourse Relations Discourse Relations in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Webber et al., 2019), as seen in ex. (1), are represented by two arguments, labeled either *Arg1* or *Arg2*, the discourse connective (in case of an explicit relation) and finally the relation sense(s) between the two, in this case both Temporal.Asynchronous.Succession and Contingency.Cause.Reason.

(1) BankAmerica climbed 1 3/4 to 30 (*Arg1*) **after** *PaineWebber boosted its investment opinion on the stock to its highest rating* (*Arg2*).

These relations are called *shallow* discourse relations since, contrary to the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988; Carlson et al., 2002), they do not recursively build a tree. The PDTB organizes their sense taxonomy, of which examples can be seen in Table 3, into three levels, with the last one denoting the direction of the relation. The PDTB annotation scheme has additionally been shown to be portable to other languages (Zeyrek et al., 2018; Long et al., 2020).

QASRL: Back in Warsaw that year, Chopin **heard** Niccolò Paganini **play** the violin, and **composed** a set of variations, Souvenir de Paganini.
What did someone **hear**? Niccolò Paganini play the violin When did someone **compose** something? that year
QAMR: Actor and television host Gary Collins died yesterday at age 74.
What kind of host was Collins? television How old was Gary Collins? 74

Table 2: Examples of QASRL and QAMR annotations.

Semantic QA Approaches Using QA structures to represent semantic propositions has been proposed as a way to generate "soft" annotations, where the resulting representation is formulated using natural language, which is shown to be more intuitive for untrained annotators (He et al., 2015). This allows much quicker, more large-scale annotation processes (FitzGerald et al., 2018) and when used in a more controlled crowd-sourcing setup, can produce high-coverage quality annotations (Roit et al., 2020).

As displayed in Table 2, both QASRL and QAMR collect a set of QA pairs, each representing a single proposition, for a sentence. In QASRL the main target is a predicate, which is emphasized by replacing all content words in the question besides the predicate with a placeholder. The answer constitutes a span of the sentence. The annotation process itself for QASRL is very controlled, by suggesting questions created with a finite-state automaton. QAMR, on the other hand, allows to freely ask all kinds of questions about all types of content words in a sentence.

QA Approaches for Discourse The relation between discourse structures and questioning has been pointed out by Van Kuppevelt (1995), who claims that the discourse is driven by explicit and implicit questions: a writer carries a topic forward by answering anticipated questions given the preceding context. Roberts (2012) introduces the term *Question Under Discussion* (QUD), which stands for a question that interlocuters accept in discourse and engage in finding its answer. More recently, there has been work on annotating QUDs, by asking workers to identify questions raised by the text and checking whether or not these raised questions get answered in the following discourse (Westera and Rohde, 2019; Westera et al., 2020). These QUD annotations are conceptually related to QADiscourse by representing discourse information through QAs, solicited from laymen speakers.

The main difference lies in the propositions captured: we collect questions that have an answer in the sentence, targeting specific relation types. In the QUD annotations (Westera et al., 2020) any type of question can be asked that might or might not be answered in the following discourse.

Previous Discourse Parsing Efforts Most of the recent work on models for (shallow) discourse parsing focuses on specific subtasks, for example on argument identification (Knaebel et al., 2019), or discourse sense classification (Dai and Huang, 2019; Shi and Demberg, 2019; Van Ngo et al., 2019). Full (shallow) discourse parsers tend to use a pipeline approach, for example by having separate classifiers for implicit and explicit relations (Lin et al., 2014), or by building different models for intra- vs. inter-sentence relations (Biran and McKeown, 2015). We also adopt the pipeline approach for our baseline model (Section 6), which performs both relation classification and argument identification, since our QA pairs jointly represent arguments and relations.

Previous Discourse Crowdsourcing Efforts There has been research on how to crowd-source discourse relation annotations. Kawahara et al. (2014) crowd-source Japanese discourse relations and simplify the task by reducing the tagset and extracting the argument spans automatically. A follow-up paper found that the data quality of these Japanese annotations was lacking compared to expert annotations (Kishimoto et al., 2018). Furthermore, Yung et al. (2019) even posit that it is impossible to crowdsource high quality discourse sense annotations and they suggest to re-formulate the task as a discourse connective insertion problem. This approach has previously also been used in other configurations (Rohde et al., 2016; Scholman and Demberg, 2017). Similarly to our QADiscourse approach, inserting connectives also works with soft natural language annotations, as we propose, but it simplifies the task greatly, by only annotating the connective, rather than retrieving complete discourse relations.

3 Representing Discourse Relations as QA pairs

In this section we discuss how to represent shallow *discourse relations* through QA pairs. For an overview, consider the second sentence in Table 1, and the two predicates 'decided' and 'knowing', each being part of a discourse unit. The sense of the discourse relation between these two units can be characterized by the question prefix "Despite what ...?" (see Table 3). Accordingly, the full QA pair represents the proposition asserted by this discourse relation, with the question and answer corresponding to the two discourse units. A complete QADiscourse representation for a text would thus consist of a set of such QAs, representing all propositions asserted through discourse relations.

The Discourse Relation Sense We want our questions to denote relation senses. To define the set of discourse relations covered by our approach, we derived a set of question templates that cover most discourse relations in the PDTB 3.0 (Webber et al., 2019; Prasad et al., 2008), as shown in Table 3. Each question template starts with a question prefix, which specifies the relation sense. The placeholder X is completed to capture the discourse unit referred to by the question, as in Table 1.

Few PDTB senses are not covered by our question prefixes. First, senses with pragmatic specifications like *Belief* and *SpeechAct* were collapsed into their general sense. Second, three *Expansion* senses were not included because they usually do not assert a new "informational" proposition, about which a question could be asked, but rather capture structural properties of the text. One of those is *Expansion.Conjunction*, which is one of the most frequently occurring senses in the PDTB, especially in intra-sentential VP conjunctions, where it makes up about 70% of the sense instances (Webber et al., 2016). Ex. (2) displays a discourse relation with two senses, one of which *Expansion.Conjunction*. While it is natural to come up with a question targeting the causal sense, the conjunction relation does not seem to assert any proposition about which an informational question may be asked.

(2) "Digital Equipment announced its first mainframe computers, *targeting IBM's* ***largest*** *market* and *heating up the industry's* ***biggest*** *rivalry*." (explicit: *Expansion.Conjunction*, implicit: *Contingency.Cause.Result*)

Finally, we removed *Purpose* as a (somewhat subtle) individual sense and subsumed it with our two causal questions.

Our desiderata for the question templates are as follows. First, we want the question prefixes to be applicable to as many scenarios as possible in

PDTB Sense	Question Template
Expansion.Substitution	Instead of what X?
Expansion.Disjunction	What is an alternative to X?
Expansion.Exception	Except when X?
Comparison.Concession	Despite what X?
Comparison.Contrast	What is contrasted with X?
Expansion.Level-of-detail	What is an example of X?
Comparison.Similarity	What is similar to X?
Temporal.Asynchronous	After/Before what X?
	Until/Since when X?
Temporal.Synchronous	While what X?
Contingency.Condition	In what case X?
Contingency.Neg.-cond.	Unless what X?
Expansion.Manner	In what manner X?
Contingency.Cause	What is the result of X?
	What is the reason X

Table 3: Informational PDTB senses mapped to our question templates.

1. And I also **feel** like in a capitalistic **society**, checks and balances **happen** when there **is** competition.
In what case do checks and balances happen? when there is competition in a capitalistic society
2. Whilst **staying** in the hotel, the Wikimedian group **met** two MEPs who **chose** it in-preference to dramatically more-expensive Strasbourg accommodation.
What is contrasted with the hotel? dramatically more-expensive Strasbourg accommodation
3. There were no fare hikes **announced** as both passenger and freight fares had been **increased** last month.
What is the reason there were no fare hikes announced? as both passenger and freight fares had been increased last month
What is the result of both passenger and freight fares having been increased last month? There were no fare hikes announced

Table 4: Example sentences with their corresponding Question-and-Answer pairs.

which discourse relations can occur, while at the same time ideally adhering to a one-to-one mapping of sense to question. Similarly, we avoid question templates that are too general. The WHEN-Question in QASRL (Table 2), for instance, can be used for either Temporal or Conditional relations. Here we employ more specific question prefixes to remove this ambiguity. Finally, as multiple relation senses can hold between two discourse units (Rohde et al., 2018), we similarly allow multiple QA pairs for the same two discourse units.

The Discourse Units The two sentence fragments, typically clauses, that we relate with a question are the discourse units. In determining what makes a discourse unit, we include verbal predicates, noun phrases and adverbial phrases as potential targets. This, for example, would also cover such instances: "Because of the rain ..." or "..., albeit warily". We call the corresponding verb, noun or adverb heading a discourse unit a *target*.

A question is created by choosing a question prefix, an auxiliary, if necessary, and copying words from the sentence. It can then be manually adjusted to be made grammatical. Similarly, the discourse unit making up the answer consists of words copied from the sentence and can be modified to be made grammatical. Our question and answer format thus deviates considerably from the QASRL representation. By not introducing placeholders, questions sound more natural and easy to answer compared to QASRL, while still being more controlled than the completely free form questions of QAMR. In addition, allowing for small grammatical adjustments introduces valuable flexibility which contribute to the intuitiveness of the annotations.

Relation Directionality Discourse relations are often directional. Our QA format introduces directionality by placing discourse units into either the question or answer. In some question prefixes, a single order is dictated by the question. As seen in ex. 1 of Table 4, because the question asks for the *condition*, the condition itself will always be in the answer. Another ordering pattern occurs for *symmetric* relations, meaning that the relation's assertion remains the same no matter how the arguments are placed into the question and answer, as in ex. 2 in Table 4. Finally, certain pairs of relation senses are considered *reversed*, such as for causal (reason vs. result) and some of the temporal (before vs. after) question prefixes. In this case, two QA pairs with different question prefixes can denote the same assertion when the target discourse units are reversed, as shown in ex. 3 in Table 4. These patterns of directionality impact annotation and evaluation, as would be described later on.

4 The Crowdsourcing Process

Pool of Annotators To find a suitable group of annotators we followed the Controlled Crowdsourcing Methodology of Roit et al. (2020). We first released two trial tasks, after which we selected the best performing workers. These workers then underwent two short training cycles, estimated to take about an hour each, which involved reading the task guidelines, consisting of 42 slides[1], completing 30 HITs per round and reading personal feedback after each round (preparing these feed-

[1] https://github.com/ValentinaPy/QADiscourse

backs consumed about 4 author work days). 11 workers successfully completed the training.

For collecting production annotations of the Dev and Test Sets, each sentence was annotated by 2 workers independently, followed by a third worker who adjudicated their QA pairs to produce the final set. For the Train Set, sentences were annotated by a single worker, without adjudication.

Preprocessing In preprocessing, question targets are extracted automatically using heuristics and POS tags: a sentence is segmented using punctuation and discourse connectives (from a list of connectives from the PDTB). For each segment, we treat the last verb in a consecutive span of verbs as a separate target. In case a segment does not contain a verb, but does start with a discourse connective, we choose one of the nouns (or adverbs) as target. The following illustrates our target extraction: *[Despite labor-shortage **warnings**,] [80% **aim** for first-year wage increases of under 4%;] [and 77% **say** they'd try to **replace** workers,] [if **struck**,] [or would **consider** it.]*

Annotation Tool and Procedure Using Amazon Mechanical Turk, we implemented two interfaces[2], one for the *QA generation* and one for the *QA adjudication* step.

In the *QA generation* interface, workers are shown a sentence with all target words in bold. Workers are instructed to generate one or more questions that relate two of these target words. The question is generated by first choosing a question prefix, then, if applicable, an auxiliary, then selecting one or more spans from the sentence to form the complete question, and lastly, change it to make it grammatical. Given the generated question, the next step involves answering that question by selecting span(s) from the sentence. Again, the answer can also be amended to be made grammatical.

The *QA adjudication* interface displays a sentence and all the QA pairs produced for that sentence by two annotators. For each QA pair the adjudicator is asked to either label it as *correct*, *not correct* or *correct, but not grammatical*. Duplicates and nonsensical QA pairs labeled as *not correct* are omitted from the final dataset. As a last step, the first author manually corrected all the *not grammatical* instances.

Data and Cost We sampled sentences from the Wikinews and Wikipedia sections of Large Scale QASRL (FitzGerald et al., 2018) while following their Train, Dev & Test splits. Descriptive statistics for the final dataset are shown in Table 5 and 6.

Annotating a sentence of Dev and Test yielded 2.11 QA pairs with a cost of 50.3¢ on average. For Train, the average cost was 37.1¢ for 1.72 QAs per sentence.

[2]Please refer to the appendix for all UI screenshots.

Dataset Split	Sentences	Questions
Wikinews Train	3098	4760
Wikinews Dev	669	1108
Wikinews Test	658	1498
Wikipedia Train	3277	6225
Wikipedia Dev	667	1524
Wikipedia Test	678	1498
Overall	9047	16613

Table 5: Dataset Statistics: Number of sentences containing at least 1 QA annotation and the total number of QAs collected.

QA Prefix	Count	Proportion
In what manner X?	4225	25%
What is the reason X?	3238	19%
What is the result of X?	2735	16 %
What is an example of X?	1757	11 %
After what X?	1099	7 %
While what X?	1060	6 %
In what case X?	509	3 %
Despite what X?	477	3 %
What is contrasted with X?	317	2 %
Before what X?	299	2 %
Since when X?	279	2 %
What is similar to X?	218	1 %
Until when X?	155	1 %
Instead of what X?	105	1 %
What is an alternative to X?	92	$\leq$ 1 %
Except when X?	27	$\leq$ 1 %
Unless what X?	21	$\leq$ 1%

Table 6: Counts of collected question types.

5 Dataset Evaluation

5.1 Evaluation Metrics

We aim to evaluate QA pairs, as the output of both the annotation process and the question generation and answering model, which are not the same as discourse relation triplets. There are multiple difficulties that arise when evaluating the QADiscourse setup. We allow multiple labels per proposition pair and thus need evaluation measures suitable for multi-label classification. Annotators are generating the questions and answers, which contrary to a pure categorical labelling task implies that we have to take into consideration question and answer paraphrasing and natural language generation inconsistencies. This requires us to use metrics that create alignments between sets of QAs, which means that existing discourse relation evaluation methods, such as from CoNLL-2015 (Xue et al.,

2015), are not applicable. The following metrics, which we apply for both the quality analysis of the dataset and the parser evaluation, are closely inspired by previous work on collecting semantic annotations with QA pairs (Roit et al., 2020; FitzGerald et al., 2018).

Unlabeled Question and Answer Span Detection (UQA) (F1) This metric is inspired by the question alignment metric for QASRL, which takes into account that there are many ways to phrase a question and therefore an exact match metric will be too harsh. Given a sentence and two sets of QA pairs produced for that sentence, such as gold and predicted sets, we want to match the QAs from the two sets for comparison. A QA pair is aligned with another QA pair that has the maximal *intersection over union* (IOU) ≥ 0.5 on a token-level, or else remains unaligned[3]. Since we allow multiple QA pairs for two targets, we also allow one-to-many and many-to-many alignments. As we are evaluating unlabeled relations at this point, we do not consider relation direction and therefore do not differentiate between question and answer spans.

Labeled Question and Answer Span Detection (LQA) (Accuracy) Given the previously produced alignments from UQA we check for the exact match of aligned question prefixes. For many-to-many and many-to-one alignments we count as correct if there is overlap of at least one question prefix. *Reversed* and *symmetric* question prefixes are converted to a more general label for fair comparison.

5.2 Dataset Quality

Inter-Annotator Agreement (IAA) To calculate the agreement between individual annotators we use the above metrics (UQA and LQA) for different worker-vs-worker configurations. The setup is the following: A set of 4 workers annotates the same sentences (around 60), from which we then calculate the agreement between all the possible pairs of workers. We repeat this process 3 times and show the average agreement scores in Table 7. The scores after adjudication, pertaining to the actual dataset agreement level, are produced by comparing the resulting annotation of two worker triplets, each consisting of two annotators and a separate adjudicator on the same data, averaged over 3 samples of 60 sentences each. These results show that adjudication notably improves agreement.

	UQA	LQA
Before Adjud.	76.87	56.64
After Adjud.	87.44	65.46

Table 7: IAA scores before and after adjudication.

5.3 Agreement with Expert Set

Our Expert set consists of 25 sentences annotated with QA pairs by the first author of the paper. Comparing the adjudicated crowdsourced annotations with the Expert Set yields a UQA (LQA) of **93.9** (**80**), indicating a high quality of our collected annotations. The main issue in disagreement arises from sentences that do not contain overt propositional discourse relations, where workers attempt to ask questions anyways, resulting in sometimes unnatural or overly implicit questions.

5.4 Comparison with PDTB

We crowdsourced QA annotations of 60 sentences from section 20 of the PDTB (commonly used as Train) with our QA annotation protocol. The PDTB arguments are aligned with the QA-pairs using the UQA metric, by considering *Arg1* and *Arg2* as the text spans to be aligned with the question and answer text[4], yielding **83.2** Precision, **87.5** Recall and an F1 of **85.3**.

A manual comparison of the PDTB labels with the Question Prefixes reveals that in most of the cases the senses overlap in meaning, with some exceptions on both sides. 60% of aligned annotations correspond exactly in the discourse relation sense they express. The remaining 40% of the QA-pairs belong to either of the following categories:

(1) Discourse relations that we deemed to be non-informational at the propositional level were many times still annotated with our QA pairs. Take this sentence: [...], *a Soviet-controlled regime remains in Kabul*, the refugees sit in their camps, and *the restoration of Afghan freedom seems as far off as ever*. The PDTB posits an *Exp.Conjunction* relation between the two cursive arguments, which is a relation type that we do not cover in the QA framework, yet our annotators saw an implied causal relation which they expressed with the following (sensible) QA pair: **What is the reason** *the restoration of Afghan freedom seems as far off as ever?* -

[3]The average length for tokenized questions and answers is 12.22 and 10.27 respectively.

[4]Such alignment is usually straightforward, since annotators do not add content words when producing QAs.

Frank Carlucci III was named to this telecommunications companyś board, filling the vacancy created by the death of William Sobey last May. (Contingency.Cause.Result)
After what was Frank Carlucci III named to this telecommunications company's board? the death of William Sobey last may *What is the reason* Frank Carlucci III was named to this telecommunications company's board? filling the vacancy created by the death of William Sobey last May

Table 8: Example of a QADiscourse relation which was not captured in the PDTB.

Question Prefix	Count
What is the reason/result of	23/20
In what manner	19
While/After/Before what	19
What is an example of	10
Since/Until when	5
In what case	4
Despite what	1

Table 9: Count of QADiscourse Question Prefixes of questions that could be aligned to QAMR.

a soviet-controlled regime remains in Kabul.

(2) Interestingly, we observe that some annotation decision difficulties described in the PDTB (Webber et al., 2019) are also mirrored in our collected data. One of those arising ambiguities is the difference between *Comparison.Contrast* and *Comparison.Concession*, in our case *Despite what* and *What is contrasted with*. In the manually analyzed data sample, 3 such confusions were found between the QADiscourse and the PDTB annotations.

(3) There were 15 instances of PDTB relation senses that were erroneously not annotated with an appropriate QA pair, even though a suitable Question Prefix exists, corresponding to some of the 12.5% recall misses in the comparison.

(4) On the contrary, there were 36 QA instances that capture appropriate propositions which were completely missed in the PDTB [5]. For example, in Table 8, the PDTB only mentions the causal relation, while QADiscourse found both the causal and the temporal sense:

Additionally we noticed that annotators tend to ask "What is similar to..?" questions about conjunctions, indicating that conjoined clauses seem to imply a similarity between them, while the similarity relation in the PDTB is rather used in more explicit comparison contexts. The "In what case..?" questions were sometimes used for adjuncts specifying a time or place. Overall, these comparisons show that agreement with the PDTB is good, with QADiscourse even finding additional valid relations, indicating that it is feasible to crowdsource high-quality discourse relations via QADiscourse.

5.5 Comparison with QAMR and QASRL

While commonly treated as two distinct levels of textual annotations, there are nevertheless some commonalities between shallow discourse relations and semantic roles. This interplay of discourse and semantics has also been noted by Prasad et al. (2015), who made use of clausal adjunct annotations in PropBank to enrich intra-sentential discourse annotations and vice versa. Similarly, we found that there are questions in QASRL, QAMR and QADiscourse which express kindred relations: *Manner*, *Condition*, *Causal* and *Temporal* relations could all be asked about using QASRL-like WH-Question. But then the point of question ambiguity arises: if "When" can be used to ask about conditional relations, it is more often also used to denote temporal relations. This under-specification becomes problematic when attempting to map between QAs and labels from resources such as PropBank. Therefore, despite the propositional overlap of some of the question types, QADiscourse additionally enriches and refines QASRL annotations.

Since QAMR does not restrict itself to predicate-argument relations only, we performed an analysis of whether annotators tend to ask about QADiscourse-type relations in a general QA setting. 965 sentences contain both QAMR and QADiscourse annotations, with 1505 QADiscourse pairs, of which we could align 101 (7%) to QAMR annotations, using the UQA-alignment algorithm. We conclude that QAMR and QADiscourse target mostly different propositions and relation types.

Within the 101 QADiscourse QAs that were aligned with QAMR questions (Table 9), causal and temporal relations are very common, usually expressed, as expected, by "Why" or "When" questions in QAMR. In other cases, the aligned questions express different relation senses. Notably, the QADiscourse *In what manner* relation aligns with a "How" QAMR question only once out of 19 cases. Often, it seems that QADiscourse annotators were tempted to ask a somewhat inappropriate *manner* question while the relation between the predicate and the answer corresponded to a direct semantic role (like location) rather than to a discourse

[5] The full list of these instances can be found in the appendix.

She said he "had friends in every political party ..."
QADISC.: *In what manner* did he have friends? QAMR: *Where* does he have friends? ANSWER: in every political party
... your internet access provider can still keep track of what websites you visit, websites can collect information about you and so on.
QADISC.: *What is an example of* something your internet access provider can still keep track of? QAMR: Your provider can keep track of *what?* ANSWER: what websites you visit

Table 10: Examples of interesting aligned cases between QAMR and QADiscourse.

relation (first example in Table 10). The second example in Table 10 corresponds to a case where the predicate-answer relation has two senses, a discourse sense captured by QADiscourse (*What is an example of*), as well as a semantic role ("theme"), captured by a "What" question in QAMR. These observations suggest interesting future research on integrating QADiscourse annotations with semantic role QA annotations, like QASRL and QAMR.

6 Baseline Model for QADiscourse

In this section we aim to devise a baseline discourse parser based on our proposed representation, which accepts a sentence as input and outputs QA pairs for all discourse relations in that sentence, to be trained on our collected data. Similarly to previous work on discourse parsing (Section (1)), our proposed parser is a pipeline consisting of three phases: *(i) question prefix prediction*, *(ii) question generation*, and *(iii) answer generation*.

Formally, given a sentence $X = x_0, ..., x_n$ with a set of indices I which mark target words (based on the target extraction heuristics in Section 4), we aim to produce a set of QA-pairs (Q_j, A_j) using the following pipeline:

1. *Question Prefix Prediction*: Let Ψ be the set of all Question Prefixes, each reflecting a relation sense from the list shown in Table 3. For each target word x_i, such that $i \in I$, we predict a set of possible question prefixes $P_{x_i} \subseteq \Psi$. The set $P = \bigcup_{i \in I} P_{x_i}$ is now defined to be the set of all prefixes for all targets in the sentence.

2. *Question Generation*: For every question prefix $p \in P$ and all its relevant target words $P_p = \{x_i | p \in P_{x_i}\}$, predict question bodies for one or more of the targets $Q_p^1, ..., Q_p^m$.

3. *Answer Generation*: Let a full question FQ_p^j be defined by the concatenation of the question prefix and the corresponding generated question body $FQ_p^j = \langle p, Q_p^j \rangle$. Given a sentence X and the question FQ_p^j, we aim to generate an answer A_p^j.

All in all, we can generate up to $|I| \times |\Psi|$ different QAs per sentence.

6.1 Question Prefix Prediction

In the first step of our pipeline we are given a sentence and a marked target, and we aim to predict a set of possible prefixes reflecting potential discourse senses for the relation to be predicted. We frame this task of predicting a set of prefixes as a multi-label classification task.

To represent I, the input sentence $X = x_0, ..., x_n$ is concatenated with a binary target indicator, and special tokens are placed before and after the target t_i. The output of the system is a set of question prefixes P_{x_i}.

We implement the model using BERT (Devlin et al., 2019) in its standard fine-tuning setting, except that the Softmax layer is replaced by a Sigmoid activation function to support multi-label classification. The predicted question prefixes are obtained by choosing those labels that have a logit $>= \tau = 0.3$, which was tuned on Dev to maximize UQA F1. Since the label distribution is skewed, we add weights to the positive examples for the binary cross-entropy loss.

6.2 Question Generation

Next in our pipeline, given the sentence, a question prefix and its relevant targets in the sentence, we aim to generate question bodies for one or more of the targets. To this end, we employ a Pointer Generator model (Jia and Liang, 2016) such that the input to the model is encoded as follows: $[CLS]$ $x_1, x_2 ... x_n$ $[SEP]$ p $[SEP]$, with $p \in P$ being the question prefix. Additionally, we concatenate a target indicator for all relevant targets P_p. The output is one or more question bodies Q_p separated by a delimiter token: Q_p^1 $[SEP]$ Q_p^2 $[SEP]$ ... Q_p^m.

The model then chooses whether to copy a word from the input, or to predict a word during decoding. We use the ALLENNLP (Gardner et al., 2018) implementation of COPYNET (Gu et al., 2016) and adapt it to work with BERT encoding of the input.

6.3 Answer Generation

To predict the answer given a full question, we use BERT fine-tuned on SQUAD (Rajpurkar et al., 2016).[6] We additionally fine-tune the model on

[6] https://huggingface.co/transformers/pretrained_models.html

1. This process, [...], rather than maintaining it as a network of unequal principalities, would ultimately be completed by Caesar's successor [...]	**4.** A writer since his teens, Pratchett first came to prominence with the Discworld novel [...]
Instead of what would this process [...] be completed by Caesar's successor? - rather than maintaining it as a network of unequal principalities	Since when did Pratchett a writer? - since his teens
2. Most decked vessels were mechanized, but two-thirds of the open vessels were traditional craft propelled by sails and oars.	**5.** Each segment of the search could last for several weeks before resupply in Western Australia.
What is contrasted with most decked vessels appearing mechanized? - two-thirds of the open vessels were traditional craft propelled by sails and oars	What is the reason each segment of the search could last for several weeks? - before resupply in Western Australia
3. It could hit Hawaii if it stays on its predicted path.	**6.** For Cook Island Maori , it was 29 % compared to 23 % ; for Tongans , 37 % to 29 % [...].
In what case could it hit Hawaii? - if it stays on its predicted path	What is contrasted with it For Cook Island Maori? - 23 %

Table 11: Examples of the QA output of the full pipeline: On the left column successful predictions and on the right wrong predictions (4: not grammatical but sensible, 5: non-sensical but grammatical, 6: neither).

	Dev	Test
UQA Precision	81.1	80.79
UQA Recall	84.94	86.8
UQA F1	82.98	83.69
LQA Accuracy	67.49	66.59
Prefix Accuracy	51.3	49.94

Table 12: Full pipeline performance for the QA-Model evaluated with labeled and unlabeled QA-alignment.

	Dev	Test
Question Prediction	71.9	65.9
Answer Prediction	68.9	72.3

Table 13: Accuracy of answers predicted by the question and answer prediction model, given a Gold question as input, compared to the Gold spans.

a subset of our training data (all 5004 instances where we could align the answer to a consecutive span in the sentence). Instead of predicting or copying words from the sentence, this model predicts start and end indices in the sentence.

7 Results and Discussion

After running the full pipeline, we evaluate the predicted set of QA-pairs against the gold set using the UQA and LQA metrics, described in section 5.1. Table 12 shows the results. Note that the LQA is dependent on the UQA, as it calculates the labeled accuracy only for QA pairs that could be aligned with UQA. The Prefix Accuracy measure complements LQA by evaluating the overall accuracy of predicting a correct question prefix. For this baseline model it shows that generally only half of the generated questions have a question prefix equivalent to gold, leaving room for future models to improve upon. While not comparable, Biran and McKeown (2015), for example, mention an F1 of 56.91 for predicting intra-sentential relations.

The scores in Table 13 show the results for the subsequent individual steps, given gold input, evaluated using a matching criterion of intersection over union $>= 0.5$ with the respective gold span.

We randomly selected a sample of 50 predicted QAs for a qualitative analysis. 22 instances from this sample were judged as correct, and 2 instances were correct despite not being mentioned in Gold. Examples of good predictions are shown on the left column in Table 11. The model is often able to learn when to do the auxiliary flip from clause to question format and when to change the verb form of the target. Interestingly, whenever the model was not familiar with a specific verb, it chose a similar verb in the correct form, for example 'appearing' in Ex. 2. The model is also able to form a question using non-adjacent spans of the sentence (Ex. 1). Some predictions do not appear in the dataset, but make sense nonetheless. The analysis showed 8 non-grammatical but sensible QAs (i.e. ex. 4, where the sense of the relation is still captured), 8 non-sensical but grammatical QAs (ex. 5) and 7 QAs that were neither (ex. 6). Lastly, we found 3 QAs with good questions and wrong answers.

8 Conclusion

In this work, we show that discourse relations can be represented as QA pairs. This intuitive representation enables scalable, high-quality annotation via crowdsourcing, which paves the way for learning robust parsers of informational discourse QA pairs. In future work, we plan to extend the annotation process to also cover inter-sentential relations.

Acknowledgments

We would like to thank Amit Moryossef for his help with the implementation of the frontend, and Julian Michael, Gabriel Stanovsky and the anonymous reviewers for their feedback and suggestions. This work was supported in part by grants from Intel Labs, Facebook, the Israel Science Foundation grant 1951/17 and the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1) and by an ERC-StG grant #677352 and an ISF grant #1739/26.

References

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

Or Biran and Kathleen McKeown. 2015. Pdtb discourse parsing as a tagging task: The two taggers approach. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 96–104.

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

Zeyu Dai and Ruihong Huang. 2019. A regularization approach for incorporating event knowledge and coreference relations into neural discourse parsing. 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 2967–2978.

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.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6.

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

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640.

Hangfeng He, Qiang Ning, and Dan Roth. 2020. Quase: Question-answer driven sentence encoding. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

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.

R. Jia and P. Liang. 2016. Data recombination for neural semantic parsing. In *Association for Computational Linguistics (ACL)*.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 269–278.

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2018. Improving crowdsourcing-based annotation of japanese discourse relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

René Knaebel, Manfred Stede, and Sebastian Stober. 2019. Window-based neural tagging for shallow discourse argument labeling. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 768–777.

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.

Junyi Jessy Li, Kapil Thadani, and Amanda Stent. 2016. The role of discourse units in near-extractive summarization. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–147.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Wanqiu Long, Xinyi Cai, James Reid, Bonnie Webber, and Deyi Xiong. 2020. Shallow discourse annotation for chinese ted talks. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1025–1032.

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.

Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. 2018. Crowdsourcing question-answer meaning representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 560–568.

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 *LREC*. Citeseer.

Rashmi Prasad, Bonnie Webber, Alan Lee, Sameer Pradhan, and Aravind Joshi. 2015. Bridging sentential and discourse-level semantics through clausal adjuncts. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 64–69.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

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

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher NL Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the blanks in understanding discourse adverbials: Consistency, conflict, and context-dependence in a crowdsourced elicitation task. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58.

Hannah Rohde, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2267.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Crowdsourcing a high-quality gold standard for qa-srl. In *ACL 2020 Proceedings, forthcoming*. Association for Computational Linguistics.

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. 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 5794–5800.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, pages 109–147.

Linh Van Ngo, Khoat Than, Thien Huu Nguyen, et al. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207.

Bonnie Webber, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 22–31.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual.

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.

Matthijs Westera and Hannah Rohde. 2019. Asking between the lines: Elicitation of evoked questions in text. In *Proceedings of the Amsterdam Colloquium*.

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

Frances Yung, Merel CJ Scholman, and Vera Demberg. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. In *The 13th Linguistic Annotation Workshop*, page 16.

Deniz Zeyrek, Amalia Mendes, and Murathan Kurfalı. 2018. Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC2018)*.

A Appendices

A.1 Reproducibility information

The code to reproduce the QADiscourse model can be found at https://github.com/ValentinaPy/QADiscourse.

Calculating the weights for the Loss of the Prefix Classifier Each question prefix label is weighted by subtracting the label count from the total count of training instances and dividing it by the label count: $weight_x = total\_instances - count_x/(count_x + 1e - 5)$.

A.2 Annotation Details

The number of examples and the details of the splits are mentioned in the paper. The data collection process has also been described in the main body of the paper. Here we add a more detailed description of the Target Extraction Algorithm and screenshots of the annotation interfaces.

Target Extraction Algorithm In order to extract targets we use the following heuristics: We split the sentence on the following punctuation: ";" ";" ":". This provides an initial incomplete segmentation of clauses and subordinate clauses. We will try to find at least one target in each segment.

We then split the resulting text spans from 1. using a set of discourse connectives. We had to remove the most ambiguous connectives from the list, whose tokens might also have other syntactic functions, for example "so, as, to, about", etc.

We then check the POS tags of the resulting spans and treat each consecutive span of verbs as a target, with the last verb in the consecutive span being the target. In order to not treat cases such as "is also studying" as separate targets, we treat "V ADV V" also as one consecutive span. In case there is no verb in a given span, we chose one of the nouns as the target, but only if the span starts with a discourse connective. This condition allows us to not include nouns as targets that are simply part of enumerations, while at the same time it helps include eventive nouns, see b) for an example. To improve precision (by 0.02) we also excluded the following verbs "said, according, spoke".

With these heuristics we achieve a Recall of 98.4 and a Precision of 57.4 compared with the discourse relations in Sec. 22 of the PDTB.

Cost Details The basic cost for each sentence was 18¢, with a bonus of 3¢ for creating a second QA pair and then a bonus of 4¢ for every additional QA pair after the first two. Adjudication was rewarded with 10¢ per sentence. On average 50.3¢ were spent per sentence of Dev and Test, with an average of 2.11 QA pairs per sentence. For Train the average cost per sentence is about 37.1¢, with an average of 1.72 QAs.

Annotation Interfaces The following screenshots display the Data Collection and Adjudication interfaces.

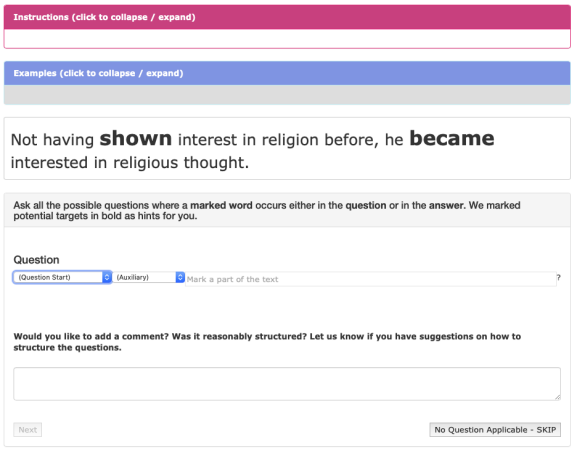

Figure 1: Interface for the Question Generation step.

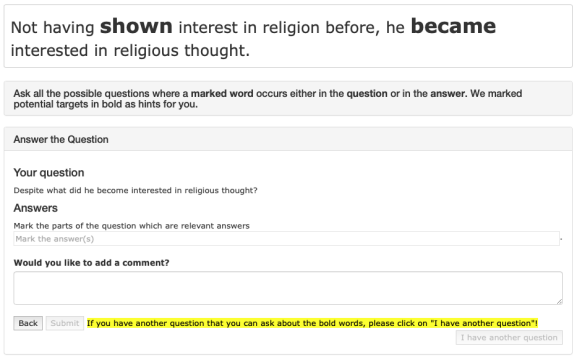

Figure 2: Interface for the Answer Generation step.

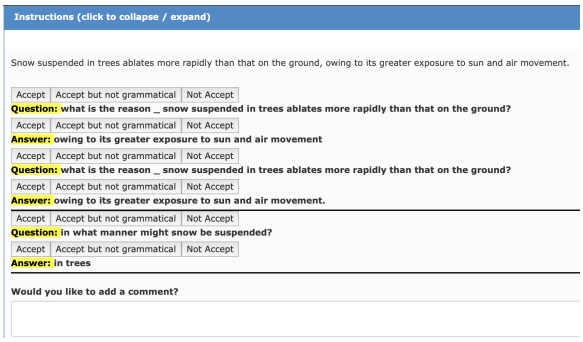

Figure 3: Interface for the Adjudication step.

A.3 Data Examples

Sentence	Question	Answer
An inquest found he'd committed suicide, but some dispute this and believe it was an accident .	Instead of what do some believe it was an accident?	suicide
On Sunday, in a video posted on YouTube, Anonymous announced their intentions saying, "From the time you have received this message, our attack protocol has past been executed and your downfall is underway."	Since when has our attack protocol past been executed and your downfall is underway?	From the time you have received this message
It is unclear why this diet works.	Despite what does this diet work?	Being unclear why
It's a downgraded budget from a downgraded Chancellor [...] Debt is higher in every year of this Parliament than he forecast at the last Budget.	What is similar to it's a downgraded budget?	It's a downgraded Chancellor
According to Pakistani Rangers, the firing from India was unprovoked in both Sunday and Wednesday incidents; Punjab Rangers in the first incident, and Chenab Rangers in the second incident, retaliated with intention to stop the firing.	What is the reason punjab Rangers and Chenab Rangers retaliated?	with intention to stop the firing
The vessel split in two and is leaking fuel oil .	After what did the vessel leak fuel oil?	The vessel split in two
In contrast to the predictions of the Met Office, the Environment Agency have said that floods could remain in some areas of England until March, and that up to 3,000 homes in the Thames Valley could be flooded over the weekend.	What is contrasted with the predictions of the Met Office?	the Environment Agency have said that floods could remain in some areas of England until March, and that up to 3,000 homes in the Thames Valley could be flooded over the weekend

Table 14: Examples for QA pairs that were annotated in the dataset.

Sentence	Question	Answer
Standard addition can be applied to most analytical techniques and is used instead of a calibration curve to solve the matrix effect problem.	Instead of what is standard addition used?	a calibration curve
State officials therefore share the same interests as owners of capital and are linked to them through a wide array of social, economic, and political ties.	What is similar to state officials share the same interests as owners of capital?	are linked to them through a wide array of social, economic, and political ties
Recently, this field is rapidly progressing because of the rapid development of the computer and camera industries.	What is the reason this field is rapidly progressing?	Because of the rapid development of the computer and camera industries
Civilization was the product of the Agricultural Neolithic Revolution; as H. G. Wells put it, "civilization was the agricultural surplus."	In what manner was civilization the product of the Agricultural Neolithic Revolution?	civilization was the agricultural surplus
The portrait shows such ruthlessness in Innocent's expression that some in the Vatican feared that Velázquez would meet with the Pope's displeasure, but Innocent was well pleased with the work, hanging it in his official visitor's waiting room.	Despite what was Innocent well pleased with The work?	The portrait shows such ruthlessness in Innocent's expression that some in the Vatican feared that Velázquez would meet with the Pope's displeasure
All tropical cyclones lose strength once they make landfall.	After what do tropical cyclones lose strength?	once they make landfall
The investigation, led by former Dutch General Patrick Cammaert , is separate from the investigation led by the UN's Human Rights Council.	What is contrasted with the investigation, led by former Dutch General Patrick Cammaert?	the investigation led by the UN's Human Rights Council

Table 15: Examples for QA pairs that were predicted with the full pipeline.

Sentence	Question	Answer	PDTB senses
It was "the Soviets' Vietnam." The Kabul regime would fall.	After what would the Kabul regime fall?	after the Soviets' Vietnam	Expansion.Conjunction
Eight months after Gen. Boris Gromov walked across the bridge into the U.S.S.R., a Soviet-controlled regime remains in Kabul, the refugees sit in their camps, and the restoration of Afghan freedom seems as far off as ever.	What is the reason the restoration of Afghan freedom seems as far off as ever?	a Soviet-controlled regime remains in Kabul	Temporal.Asynchronous.Succession, Expansion.Conjunction
Soviet leaders said they would support their Kabul clients by all means necessary–and did.	In what manner would soviet leaders support their Kabul clients?	soviet leaders said they would support their kabul clients by all means necessary	Expansion.Conjunction
Soviet leaders said they would support their Kabul clients by all means necessary–and did.	After what did Soviet leaders support their Kabul clients by all means necessary?	after soviet leaders said they would	Expansion.Conjunction
With the February 1987 U.N. accords "relating to Afghanistan," the Soviet Union got everything it needed to consolidate permanent control.	In what manner did the Soviet Union get everything it needed to consolidate permanent control?	with the February 1987 u.n. accords "relating to Afghanistan,"	Contingency.Cause.Reason, Contingency.Purpose.Arg2-as-goal
The terms of the Geneva accords leave Moscow free to provide its clients in Kabul with assistance of any kind–including the return of Soviet ground forces–while requiring the U.S. and Pakistan to cut off aid.	What is the result of the terms of the Geneva accords?	leaving Moscow free to provide its clients in Kabul with assistance of any kind while requiring the U.S. and Pakistan to cut off aid	Temporal.Synchronous, Comparison.Contrast
The only fly in the Soviet ointment was the last-minute addition of a unilateral American caveat, that U.S. aid to the resistance would continue as long as Soviet aid to Kabul did.	What is the reason for the only fly in the Soviet ointment?	the last-minute addition of a unilateral American caveat, that U.S. aid to the resistance would continue as long as Soviet aid to Kabul did	Expansion.Level-of-detail.Arg2-as-detail, Contingency.Condition.Arg2-as-cond, Temporal.Synchronous
But as soon as the accords were signed, American officials sharply reduced aid.	In what manner did American officials reduce aid?	American officials sharply reduced aid	Temporal.Asynchronous.Succession
Moscow claims this is all needed to protect the Kabul regime against the guerrilla resistance.	What is the reason Moscow claims this is all needed?	to protect the Kabul regime against the guerrilla resistance	Contingency.Condition.Arg2-as-cond
But this is not the entire Afghan army, and it is no longer Kabul's only military force.	What is similar to it not being the entire Afghan army?	is no longer Kabul's only military force.	Expansion.Conjunction
The deal fell through, and Kandahar remains a major regime base.	After what did Kandahar remain a major regime base?	after the deal fell through	Contingency.Cause.Result, Expansion.Conjunction
The deal fell through, and Kandahar remains a major regime base.	Since when does Kandahar remain a major regime base?	since the deal fell through	Contingency.Cause.Result, Expansion.Conjunction
The wonder is not that the resistance has failed to topple the Kabul regime, but that it continues to exist and fight at all.	Despite what is the wonder that it continues to exist and fight at all?	despite the resistance failing to topple the kabul regime	Comparison.Contrast, Expansion.Substitution.Arg2-as-subst
Last summer, in response to congressional criticism, the State Department and the CIA said they had resumed military aid to the resistance months after it was cut off; but it is not clear how much is being sent or when it will arrive.	what is the result of congressional criticism last summer?	the state department and the CIA said they had resumed military aid to the resistance	Temporal.Asynchronous.Succession, Comparison.Concession.Arg2-as-denier, Expansion.Conjunction

Table 16: Examples for additional relations expressed through QA pairs that do not appear in the PDTB, Part 1.

Sentence	Question	Answer	PDTB senses
Beyond removing a competitor, the combination should provide " synergies," said Fred Harlow, Unilab's chief financial officer.	While what should the combination provide synergies?	removing a competitor.	Expansion.Conjunction
In Los Angeles, for example, Central has had a strong market position while Unilab's presence has been less prominent, according to Mr. Harlow.	In what case has Central had a strong market position while Unilab's presence has been less prominent?	in Los Angeles	Comparison.Contrast, Temporal.Synchronous
A Daikin executive in charge of exports when the high-purity halogenated hydrocarbon was sold to the Soviets in 1986 received a suspended 10-month jail sentence.	What is the result of the high-purity halogenated hydrocarbon being sold to the Soviets in 1986?	a Daikin executive in charge of exports received a suspended 10-month jail sentence	Temporal.Synchronous
In Los Angeles, for example, Central has had a strong market position while Unilab's presence has been less prominent, according to Mr. Harlow.	in what case has central had a strong market position while Unilab's presence has been less prominent?	in Los Angeles	Comparison.Contrast, Temporal.Synchronous
Mr. Mehl noted that actual rates are almost identical on small and large-denomination CDs, but yields on CDs aimed at the individual investor are boosted by more frequent compounding.	In what manner are yields on CDs aimed at the individual investor boosted?	by more frequent compounding	Comparison.Concession.Arg2-as-denier
Judge Masaaki Yoneyama told the Osaka District Court Daikin's "responsibility is heavy because illegal exports lowered international trust in Japan." Sale of the solution in concentrated form to Communist countries is prohibited by Japanese law and by international agreement.	Except when is the solution in concentrated form sold?	except to communist countries	Contingency.Cause.Reason, EntRel
Japan has supported a larger role for the IMF in developing-country debt issues, and is an important financial resource for IMF-guided programs in developing countries.	In what case is Japan an important financial resource for imf-guided programs?	in developing countries	Expansion.Conjunction
Japan has supported a larger role for the IMF in developing-country debt issues, and is an important financial resource for IMF-guided programs in developing countries.	While what has Japan supported a larger role for the IMF in developing-country debt issues?	while it is an important financial resource for imf-guided programs in developing countries	Expansion.Conjunction
The last U.S. congressional authorization, in 1983, was a political donnybrook and carried a $6 billion housing program along with it to secure adequate votes.	What is an example of something being a political donnybrook?	the last u.s. congressional authorization, in 1983	Contingency.Purpose.Arg2-as-goal, Expansion.Conjunction
Instead, the tests will focus heavily on new blends of gasoline, which are still undeveloped but which the petroleum industry has been touting as a solution for automobile pollution that is choking urban areas.	What is the reason tests will focus heavily on new blends of gasoline?	the petroleum industry has been touting as a solution for automobile pollution	Comparison.Concession.Arg2-as-denier
While major oil companies have been experimenting with cleaner-burning gasoline blends for years, only Atlantic Richfield Co. is now marketing a lower-emission gasoline for older cars currently running on leaded fuel.	While what is Atlantic Richfield co. marketing a lower-emission gasoline for older cars currently running on leaded fuel?	while major oil companies have been experimenting with cleaner-burning gasoline blends	Comparison.Contrast

Table 17: Examples for additional relations expressed through QA pairs that do not appear in the PDTB, Part 2.

Sentence	Question	Answer	PDTB senses
Instead, a House subcommittee adopted a clean-fuels program that specifically mentions reformulated gasoline as an alternative.	What is the result of a house subcommittee adopting a clean-fuels program?	reformulated gasoline as an alternative.	Expansion.Level-of-detail.Arg2-as-detail
The Bush administration has said it will try to resurrect its plan when the House Energy and Commerce Committee takes up a comprehensive clean-air bill.	In what case will the Bush administration try to resurrect its plan?	when the house energy and commerce committee takes up a comprehensive clean-air bill	Temporal.Synchronous
That compares with per-share earnings from continuing operations of 69 cents the year earlier; including discontinued operations, per-share was 88 cents a year ago.	In what manner does that compare with per-share earnings from continuing operations of 69 cents the year earlier?	including discontinued operations, per-share was 88 cents a year ago.	Comparison.Contrast
Analysts estimate Colgate's sales of household products in the U.S. were flat for the quarter, and they estimated operating margins at only 1% to 3%	While what did analysts estimate Colgate's sales of household products in the U.S. were flat for the quarter?	they estimated operating margins at only 1% to 3%	Expansion.Conjunction
Analysts estimate Colgate's sales of household products in the U.S. were flat for the quarter, and they estimated operating margins at only 1% to 3%	After what did analysts estimate Colgate's sales of household products in the U.S. were flat?	after the quarter	Expansion.Conjunction
The programs will be written and produced by CNBC, with background and research provided by staff from U.S. News	What is similar to the programs being written by CNBC?	being produced by CNBC	Expansion.Conjunction
The programs will be written and produced by CNBC, with background and research provided by staff from U.S. News	In what manner will background and research be provided for the programs?	by staff from U.S. news	Expansion.Conjunction
The programs will be written and produced by CNBC, with background and research provided by staff from U.S. News	In what manner will the programs be written and produced?	the programs will be written and produced by CNBC, with background and research provided by staff from U.S. news	Expansion.Conjunction
Frank Carlucci III was named to this telecommunications company's board, filling the vacancy created by the death of William Sobey last May.	After what was Frank Carlucci III named to this telecommunications companyś board?	the death of William Sobey last may	Contingency.Cause.Result
Weyerhaeuser's pulp and paper operations were up for the nine months, but full-year performance depends on the balance of operating and maintenance costs, plus pricing of certain products, the company said.	What is contrasted with full-year performance of Weyerhaeuser's pulp and paper operations?	nine month performance	Comparison.Concession.Arg2-as-denier
Weyerhaeuser's pulp and paper operations were up for the nine months, but full-year performance depends on the balance of operating and maintenance costs, plus pricing of certain products, the company said.	What is the result of Weyerhaeuser's full-year performance?	depends on the balance of operating and maintenance costs, plus pricing of certain products.	Comparison.Concession.Arg2-as-denier

Table 18: Examples for additional relations expressed through QA pairs that do not appear in the PDTB, Part 3.