ESTER: A Machine Reading Comprehension Dataset for Reasoning about Event Semantic Relations

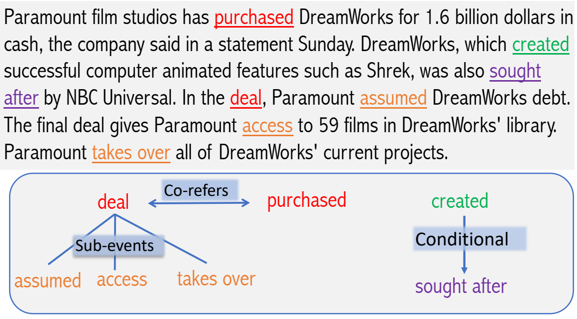
Rujun Han[1] **I-Hung Hsu**[1] **Jiao Sun**[1] **Julia Baylon**[2]
Qiang Ning[3*] **Dan Roth**[4] **Nanyun Peng**[1,2]
[1]University of Southern California [2]University of California, Los Angeles
[3]Amazon [4]University of Pennsylvania
{rujunhan,ihunghsu,jiaosun}@usc.edu; juliabaylon@ucla.edu
qning@amazon.com; danroth@seas.upenn.edu
violetpeng@cs.ucla.edu

Abstract

Understanding how events are semantically related to each other is the essence of reading comprehension. Recent event-centric reading comprehension datasets focus mostly on event arguments or temporal relations. While these tasks partially evaluate machines' ability of narrative understanding, human-like reading comprehension requires the capability to process event-based information beyond arguments and temporal reasoning. For example, to understand causality between events, we need to infer motivation or purpose; to establish event hierarchy, we need to understand the composition of events. To facilitate these tasks, we introduce *ESTER*, a comprehensive machine reading comprehension (MRC) dataset for **E**vent **S**eman**t**ic **R**elation **R**easoning. The dataset leverages natural language queries to reason about the five most common event semantic relations, provides more than 6K questions, and captures 10.1K event relation pairs. Experimental results show that the current SOTA systems achieve 22.1%, 63.3% and 83.5% for token-based exact-match (**EM**), F_1 and event-based **HIT@1** scores, which are all significantly below human performances (36.0%, 79.6%, 100% respectively), highlighting our dataset as a challenging benchmark. [1]

1 Introduction

Narratives such as stories and news articles are composed of series of events (Carey and Snodgrass, 1999; Harmon, 2012). Understanding how events are logically connected is essential for reading comprehension (Caselli and Vossen, 2017; Mostafazadeh et al., 2016b). For example, Figure 1 illustrates several pairwise relations for events in the given passage: *"the deal"* can be considered as the same event of *"Paramount purchased DreamWorks,"* forming a coreference relation; it is also a complex event that contains *"assumed debt," "gives access"* and *"takes over projects"* as its sub-events. The event *"sought after"* is facilitated by a previous event *"created features."* By capturing these event semantic relations, people can often grasp the gist of a story. Therefore, for machines to achieve human-level narrative understanding, we need to test and ensure models' capability to reason over these event relations.



Paramount film studios has purchased DreamWorks for 1.6 billion dollars in cash, the company said in a statement Sunday. DreamWorks, which created successful computer animated features such as Shrek, was also sought after by NBC Universal. In the deal, Paramount assumed DreamWorks debt. The final deal gives Paramount access to 59 films in DreamWorks' library. Paramount takes over all of DreamWorks' current projects.

Figure 1: A graph illustration of event semantic relations in narratives. We use trigger words to represent events in this graph.

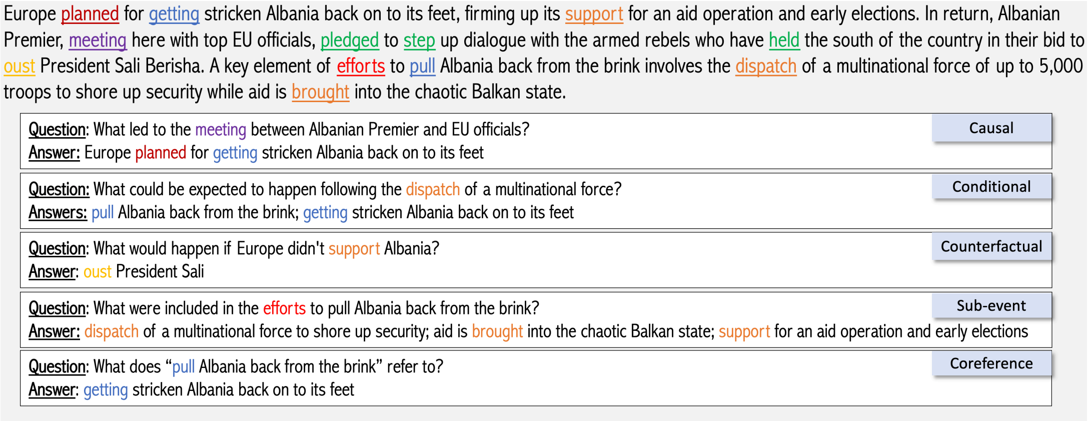
In this work, we study five types of **event semantic relations**: CAUSAL, SUB-EVENT, CO-REFERENCE, CONDITIONAL and COUNTERFACTUAL, and propose to use natural language questions to reason about **event semantic relations**. Figure 2 shows example question-answer pairs for each relation type.

Although previous works studied some subset of these relations such as SUB-EVENT (Glavaš et al., 2014; Yao et al., 2020), CAUSAL and CONDITIONAL (Mirza et al., 2014; Mirza and Tonelli, 2014; O'Gorman et al., 2016), most of them adopted pairwise relation extraction (RE) formulation by constructing (event, event, relation) triplets and predicting the relation for the pair of events. Event relations of RE formulation are rigidly defined as class labels based on expert knowledge, which could suffer from relatively low

[*] Part of the work was done while the author was at the Allen Institute for AI.

[1]Data, models and reproduction code are available here: https://github.com/PlusLabNLP/ESTER.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7543–7559
November 7–11, 2021. ©2021 Association for Computational Linguistics



Europe planned for getting stricken Albania back on to its feet, firming up its support for an aid operation and early elections. In return, Albanian Premier, meeting here with top EU officials, pledged to step up dialogue with the armed rebels who have held the south of the country in their bid to oust President Sali Berisha. A key element of efforts to pull Albania back from the brink involves the dispatch of a multinational force of up to 5,000 troops to shore up security while aid is brought into the chaotic Balkan state.

Question: What led to the meeting between Albanian Premier and EU officials?
Answer: Europe planned for getting stricken Albania back on to its feet

Question: What could be expected to happen following the dispatch of a multinational force?
Answers: pull Albania back from the brink; getting stricken Albania back on to its feet

Question: What would happen if Europe didn't support Albania?
Answer: oust President Sali

Question: What were included in the efforts to pull Albania back from the brink?
Answer: dispatch of a multinational force to shore up security; aid is brought into the chaotic Balkan state; support for an aid operation and early elections

Question: What does "pull Albania back from the brink" refer to?
Answer: getting stricken Albania back on to its feet

Figure 2: Examples of event annotations and 5 types of QAs in our dataset. Not all events are annotated for clarity purpose. Different colors are used for better visualization.

inter-annotator agreements (Glavaš et al., 2014; O'Gorman et al., 2016) and may not be the most natural way to exploit the semantic connections between relations and events in their context.

We instead propose to reason about *event semantic relations* as a *reading comprehension / question answering* task. Natural language queries ease the annotation efforts in the RE formulation by supplementing expert-defined relations with textual prompts. When querying CAUSAL relations, we can ask *"what causes / leads to Event A?"* or *"why does A happen?"*; when reasoning SUB-EVENT relation, we can ask *"what are included in Event B?"* or *"What does B entail?"* etc. *"lead to," "included"* and *"entail,"* as textual cues, can help models better understand which relation is being queried.

Our question-answering task also poses unique challenges for reasoning *event semantic relations*. First, the correct answers can be completely different with slight changes of queries. In Figure 2, if we modify the third question to be *"What would happen if Europe supported Albania?"* then *"oust President Sali"* becomes an invalid answer. This challenge allows us to test whether models possess robust reasoning skills or simply conduct pattern matching. Second, answers must be in the form of complete and meaningful text spans. For the COUNTERFACTUAL example in Figure 2, a random text span *"President Sali Berisha"* is not a meaningful answer while a shortened answer *"oust"* is not complete. To get correct answers, models need to detect both event triggers and their event arguments. Finally, there could be multiple valid events in a passage that can answer a question, and a good system should be able to identify different valid answers simultaneously as in the SUB-EVENT QA of Figure 2. These challenges make our task more difficult than the classification tasks in RE.

A few noticeable event-centric MRC datasets have been proposed recently. TORQUE (Ning et al., 2020b) and MCTACO (Zhou et al., 2019) are two recent MRC datasets that study event temporal relations. However, knowing only the temporal aspect of events could not solve many important event semantic relations. For example, in Figure 1, to understand that *"assumed debt," "gives access"* and *"takes over projects"* are sub-events of *"the deal,"* a model not only needs to know that all these four events have overlapped time intervals but also share the same associated participants for *"the deal"* to contain the other three.

We summarize our contributions below.

1. We introduce *ESTER*, the first comprehensive MRC / QA dataset for the five proposed **event semantic relations** by adopting natural language questions and requiring complete event spans in the passage as answers.
2. By proposing a generative QA task that models all five relations jointly and comparing it with traditional extractive QA task, we provide insights on how these event semantic relations interplay for MRC.
3. Our experimental results reveal SOTA models' deficiencies in our target tasks, which demonstrates that *ESTER* is a challenging dataset that can facilitate future research in MRC for **event semantic relations**.

2 Definitions

Composing event-centric questions and answers requires identifications of both events and their relations. In this section, we describe our definitions of events and five **event semantic relations**.

2.1 Events

Adopting the general guideline of ACE (2005), we define an event as a trigger word and its arguments (subject, object, time and location). An event trigger is a word that most clearly describes the event's occurrence, and it is often a verb or noun that evokes the action or the status of the target event (Pustejovsky et al., 2003). Later event-centric reasoning work mostly uses this trigger definition, e.g., TE3 (UzZaman et al., 2013), HiEve (Glavaš et al., 2014), RED (O'Gorman et al., 2016) and TORQUE (Ning et al., 2020b).

While event triggers must exist in the context, some event arguments need to be inferred by annotators. In Figure 2, for example, *"getting"* is an event trigger and its subject, object and location are *"Europe," "Albania"* and *"Europe"* respectively. The event's time can be inferred to be approximately the document writing time. To ensure event-centric reasoning, we require all questions and answers to include a trigger. Annotators are allowed to use any event arguments including those inferred to make questions natural and descriptive; whereas for answers, they need to identify complete and meaningful text spans in the passage.

2.2 Event Semantic Relations

Next, we discuss the definitions of the five types of event semantic relations in our dataset, most of which are consistent with previous studies. For example, CAUSAL and CONDITIONAL have been studied in Wolff (2007); Do et al. (2011); Mirza and Tonelli (2014); Mirza et al. (2014). SUB-EVENT and CO-REFERENCE were studied in Glavaš et al. (2014); O'Gorman et al. (2016). Cosmos QA (Huang et al., 2019) has a small amount of COUNTERFACTUAL questions, but it is not an event-centric dataset. The examples we use below are all presented in Figure 2.

Causal: A pair of events (e_i, e_j) exhibits a CAUSAL relation **if e_i happens then e_j will definitely happen** according to the given passage. For example, the passage explicitly says that the *"meeting"* happens *"in return"* for *"Europe planned for getting stricken Albanian back."* Therefore, the CAUSAL relation in the example can be established because if *"Europe planned for getting stricken Albanian back"* happens, the *"meeting"* will definitely happen in this context.

Conditional: A pair of events (e_i, e_j) exhibits a CONDITIONAL relation **if e_i facilitates, but may not necessarily leads to e_j** according to the given passage. For example, the expectation of *"the dispatch of a multinational force"* is to *"pull Albania back from the brink"*; in other words, the former event can help but does not guarantee the occurrence of the latter one. Therefore, the relation between this pair of events is CONDITIONAL.

Counterfactual: **e_j may happen if e_i does not happen**; in other words, if the negation of e_i facilitates e_j, then (e_i, e_j) has a COUNTERFACTUAL relation. In our example, if *"Europe didn't support Albania,"* which is a negation of what happens in the passage, then *"oust President Sali"* by the *"armed rebels"* would likely happen.

Sub-event: There is a semantic hierarchy where **a complex event e_k consists of a set of sub-events** $\{e_{k,1}, ..., e_{k,j}, ..., e_{k,n}\}$. In SUB-EVENT relations, we require not only $e_{k,j}$'s trigger word to be semantically contained in e_k's trigger, but also the arguments of $e_{k,j}$ are either identical or contained in the associated arguments of e_k. For example, considering the complex event *"efforts to pull Albania back,"* and its sub-event *"aid is brought into the chaotic Balkan state"*, the trigger "brought" is a part of the "efforts." Both subjects are "Europe," both objects / locations are "Albania" or "Balkan state" and their time can be inferred to be (nearly) identical in the passage. Note that this definition is similar to the event hierarchical structure definition in RED, but stricter than the "Spatial-temporal containment" definition in HiEve.

Coreference: **e_i co-refers to e_j when two events are mutually replaceable**. This requires 1) their event triggers are semantically the same and 2) their event arguments are identical. In our example, the event triggers in the question *"pull" (back from the brink)* and in the answer *"getting" (back on to its feet)* are semantically the same. They also share the same subject - Europe, and object - Albania. Their time and location can be inferred from the passage to be the same. Therefore, these two events form a CO-REFERENCE relation.

3 Related Work

We briefly survey related work in this section in order to provide broader background over the two key components of *ESTER*: 1) event semantic relations and 2) event-centric reading comprehension.

3.1 Event Semantic Relations

Event semantic relations have been studied before and most of them leverage relation extraction formulation for annotations. Causality is one of the widely studied event semantic relations. Mirza and Tonelli (2014); Mirza et al. (2014) follow the CAUSE, ENABLE and PREVENT schema proposed by Wolff (2007) where the first two relations align with our definitions in *ESTER*. Do et al. (2011) adopted a minimally supervised method and measure event causality based on pointwise mutual information of predicates and arguments, which resulted in denser annotations than previous works.

HiEve (Glavaš et al., 2014) defines pairwise SUB-EVENT relation as spatiotemporal containment, which is less rigorous than our definitions where we require containment for all event arguments (subject, object, time and location). Our definition of CO-REFERENCE is nearly identical as HiEve where two co-referred events denote the same real-world events. Yao et al. (2020) utilized a weakly-supervised method to extract large scale SUB-EVENT pairs, but the extracting rules can result in noisy relations.

RED (O'Gorman et al., 2016) proposed to annotate event temporal and semantic relations (CAUSAL, SUB-EVENT) jointly. However, due to the complexity of the annotation schema, the data available for semantic relations are relatively sparse. Mostafazadeh et al. (2016b) and Caselli and Vossen (2017) annotate both event temporal and semantic relations in ROCStories (Mostafazadeh et al., 2016a) and Event StoryLine Corpus (Caselli and Vossen, 2017) respectively. *ESTER* differs from these works by disentangling temporal from other semantic relations and focusing on MRC to capture five proposed event semantic relations.

3.2 Event-centric MRC

Datasets leveraging natural language queries for event-centric machine reading comprehension have been proposed recently (Zhou et al., 2019; Ning et al., 2020b). However, they focus on event temporal commonsense, whereas *ESTER* studies other event semantic relations. Du and Cardie (2020) and Liu et al. (2020) reformulate event extraction data as QA tasks to detect event triggers and arguments in a short passage. However, they did not propose new data, and knowing event triggers and arguments are merely a sub-task in *ESTER*, which require both event detection and relation understanding.

4 Data Collection

In this section, we show our data collection procedure and describe the details of our approach to control annotation quality, including qualification exams and steps to validate and train workers.

4.1 Passage Preparation

Passages are selected from news articles in TempEval3 (TE3) workshop (UzZaman et al., 2013) with initial event triggers provided. We extracted 3-4 continuous sentences that contain at least 7 event triggers. Our choice of the number of sentences is based on previous studies that hierarchical relations such as SUB-EVENT and CO-REFERENCE are likely to span over multiple sentences, but the majority of them are contained within 3-4 sentences (Glavaš et al., 2014; O'Gorman et al., 2016).

4.2 Main Procedure

We use Figure 2 to illustrate our main data collection procedure, which consists of two components: event selection and QA annotations. The actual interface can be found in the appendix.

Event Selections. Annotators are presented with a passage and initial event trigger annotations. They are allowed to modify event trigger selections per our definition in Section 2 by highlighting words. These correspond to the highlighted words in the passage of Figure 2. Our focus is not event extraction, and thus we do not require workers to identify all triggers as some of them are not used in their QAs. Rather, the event selection serves as a warm-up step for the following QA annotations by 1) helping workers locate where desirable events are and 2) ensuring that all the annotated question-answer pairs include events in the passage so that their QAs reason about event relations.

QA Annotations. As the five questions in Figure 2 show, users must ask natural language questions that contain a highlighted event trigger. In order to make questions natural, we allow workers to use different textual forms of an event trigger in the questions, such as "teach" v.s. "taught" and "meeting" v.s. "meet." After writing a question, users need to pick the event semantic type (the blue boxes in Figure 2) that they reason about, and then select the corresponding answer spans from the passage. If there are multiple answers, we instruct users to select *all* of them. All answers must include an exact highlighted event trigger, and we prohibit answers with more than 12 words to ensure conciseness. We pay $7.5 for an assignment where annotators need to ask at least five questions using two passages.

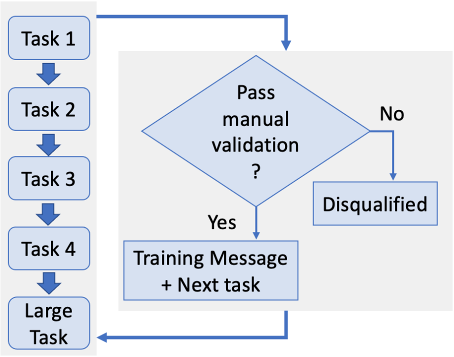


Figure 3: An illustration of our quality control, worker validation and training process.

4.3 Quality Control

Qualification. The initial worker qualification was conducted via an examination in the format of multiple-choice questions hosted by CROWDAQ platform (Ning et al., 2020a). We created a set of questions where a passage and a pair of QA are provided, and workers need to judge the correct type of this QA from six choices, including the five defined event semantic relations, plus an invalid option[2]. This examination intends to test workers' skills to 1) distinguish valid QAs from invalid ones based on our definitions; 2) judge the differences for the five proposed **event semantic relations**.

We recruit workers via Amazon Mechanical Turk with basic qualifications including: 1) at least 1K HITs[3] approved; 2) at least 97% approval rate. A single qualification exam consists of 10 multiple-choice questions. Participants are given 3 attempts to pass with a >= 0.6 score. We found this qualification examination effectively reduces the rate of spammers to nearly 0%.

Worker Validation and Training. Since the real task is much more challenging than the qualification exams, we adopted a meticulous five-stage worker validation and training process to ensure data quality. As Figure 3 shows, for workers who passed the qualification exams, we repeat the validation and training steps four times until workers reach the final large tasks.

In each validation and training step, two of our co-authors independently judge workers' annotations to determine 1) whether a provided QA pair is valid per our definitions and 2) whether the answers provided are complete. Typically, we disqualify workers whose QA validity rate falls below 90%. Exceptions are given upon careful examination and reviewer discussion. For workers who pass a manual validation, we write a training message correcting all errors they made and invite them to the next task. We also add missing answers as a part of the validation process and reserved the validated annotations as our evaluation data.

There are 1, 2, 3, 10, and 25 HITs in Task 1-4 and Large Task respectively. For Task 1-3, we validate all QAs, and for Task 4, we randomly select 20% questions per worker to validate. In order to work on the final large task, a worker needs to maintain an average QA validity rate higher than 90%. We further request one co-author to validate all questions with passages overlapped with the validated data above. This ensures that there are no passage overlaps between the training and evaluation data. All author validated data comprise our final evaluation data in the experiments.

5 Data Analysis

Our passage preparation (Section 4.1) produces 4.3K passages in total with 1887 of them randomly selected and annotated. We collect 6018 questions from 70 workers using Amazon Mechanical Turk and 1471 of them fully validated by co-authors as the evaluation set. We further split our evaluation data into dev and test sets based on passages. The remaining data are used as the training set. A summary of data statistics is shown in Table 1.

	Train	Dev	Test
# of Passages	1492	108	287
# of Questions - Overall	4547	301	1170
- CAUSAL	2047	118	431
- CONDITIONAL	928	58	289
- COUNTERFACTUAL	294	28	106
- SUB-EVENT	678	59	204
- CO-REFERENCE	600	38	140

Table 1: Passages and questions (overall + type breakdown) statistics for different data splits.

5.1 Type Distribution

As we can observe in Table 1 and Figure 9 - 10 in the appendix, *ESTER* consists of 64.2% CAUSAL and CONDITIONAL questions. In Figure 4, we further show the type disagreements using data validated by two co-authors. The rows indicate workers' original types and the columns are the majority votes between the annotators and co-authors. As we can observe, the matrix is dominated by diagonal entries. Some noticeable disagreements are 1) between CAUSAL and CONDITIONAL where people have different opinions on

[2] A full list of QA validity can be found in Appendix B.
[3] HIT is an assignment unit on Amazon Mechanical Turk.

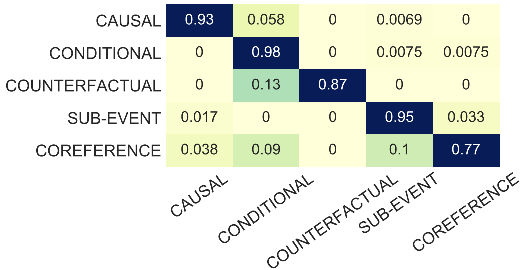


Figure 4: Type confusion matrix between workers' original annotations and the majority votes after co-authors' validation. Rows are annotators' types whereas columns are the majority votes.

the degree of causality between events; 2) between COUNTERFACTUAL and CONDITIONAL as some COUNTERFACTUAL questions, with double negations[4], are merely CONDITIONAL; 3) between CO-REFERENCE and SUB-EVENT where annotated co-referred events do not have identical event arguments according to co-authors' judgements. These results align with previous studies that some event semantic relations are inherently hard to distinguish (Glavaš et al., 2014; O'Gorman et al., 2016).

Type Agreements. The inter-annotator-agreement (IAA) score is 85.71% when calculated using pair-wise micro F1 scores, and is 0.794 per Fleiss's κ[5]. The IAA scores are calculated using the same data reported in Figure 4. The high IAA scores demonstrate strong alignments between annotators and co-authors in judging event semantic relations.

5.2 Other Statistics

We show n-grams in questions and the number of answers below. More analysis on tokens and worker distributions can be found in Appendix-E F.

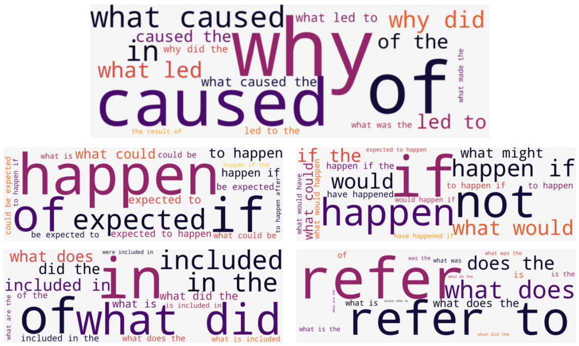


Figure 5: Most frequent n-grams in questions for each semantic type. First row: CAUSAL; second row: CONDITIONAL + COUNTERFACTUAL; third row: SUB-EVENT + CO-REFERENCE.

Frequent N-grams in Questions. Figure 5 illustrates the most frequent unigram, bigram and tri-grams in each type of questions after removing non-informative stop-words. These n-grams can be considered as semantic cues in the questions to reason about particular semantic relations. For example, 'why' and 'what caused' imply strong causality; 'included' indicates containment of events; 'not' in COUNTERFACTUAL indicates negation of events.

Number of Answers. Table 2 shows the average number of answers for each semantic type. SUB-EVENT contains the most answers, which aligns with our intuition that a complex event in the passage often contains multiple sub-events. The evaluation sets contain about 0.5 answers more than the training set as co-authors added the missing answers in the validation process. Considering each unique question and answer as an event, *ESTER* captures 10.1K event pairs, which are larger than previous RE datasets such as RED and HiEve.

Semantic Types	Train	Dev	Test
CAUSAL	1.3	1.5	1.9
CONDITIONAL	1.3	1.9	2.0
COUNTERFACTUAL	1.2	1.3	1.7
SUB-EVENT	3.0	3.6	3.1
CO-REFERENCE	1.2	1.2	1.6

Table 2: Average number of answers by semantic types.

6 Experimental setup

We design experiments to provide benchmark performances and understand learning challenges to facilitate future research on *ESTER*. We formulate our QA task as a conditional answer generation problem. This choice is inspired by recent works such as UnifiedQA (Khashabi et al., 2020) that achieve impressive outcomes by integrating various QA tasks (extractive, abstractive and multiple-choice) as a single generative QA pre-training task. Li et al. (2021) and Paolini et al. (2021) also show that by reformulating original extractive tasks as generation tasks, it enables models to better exploit semantic relations between context and labels as well as the dependencies between different outputs. To better demonstrate the benefits of the proposed generative QA task, we compare it with a traditional extractive QA task. We introduce our experimental design and evaluation metrics subsequently.

6.1 Generative QA

Given a question q_i and a passage $P_i = \{x_1, x_2, ...x_j, ...x_n\}$ where x_j represents a token in the passage, the answer generation task requires the model to generate natural language an-

[4] Double negated questions have the form of *"what will not happen if Event A does not happen"*

[5] 0.794 implies substantial agreement(Landis and Koch, 1977). The detailed calculation can be found in appendix C.

	Dev			Test		
	F_1^T	HIT@1	EM	F_1^T	HIT@1	EM
Generative Zero-shot: T5-base	18.0	55.8	0.0	21.1	61.0	0.0
Generative Zero-shot: UnifiedQA-base	49.0	61.5	10.6	46.5	61.5	7.1
Generative Zero-shot: UnifiedQA-large	51.1	69.4	14.3	48.7	66.5	9.7
Generative Fine-tune: BART-base	53.1(±0.4)	66.9(±1.7)	14.1(±1.0)	53.3(±0.8)	68.1(±1.2)	15.1(±0.7)
Generative Fine-tune: BART-large	57.2(±1.0)	72.1(±1.4)	15.1(±2.1)	56.1(±1.0)	71.5(±2.2)	15.2(±0.9)
Generative Fine-tune: T5-base	63.2(±1.1)	80.8(±1.7)	22.1(±0.9)	58.5(±0.7)	76.2(±1.0)	20.5(±0.9)
Generative Fine-tune: UnifiedQA-base	64.6(±0.4)	82.0(±0.4)	23.8(±1.0)	59.3(±0.2)	78.1(±0.4)	20.6(±0.5)
Generative Fine-tune: UnifiedQA-large	66.8(±0.2)	**87.2(±0.3)**	**24.4(±0.3)**	63.3(±0.8)	**83.5(±0.7)**	**22.1(±0.4)**
Extractive Fine-tune: RoBERTa-large	**68.8(±0.7)**	66.7(±1.1)	16.7(±0.2)	**66.1(±0.2)**	63.8(±1.6)	15.9(±0.5)
Human Baseline	-	-	-	79.6	100	36.0

Table 3: Experimental results for answer generation. All numbers are 3-seed average with standard deviation reported, except for human baseline and zero-shot performances. All models refer to the generative QA task except for RoBERTa-large, which we use for the extractive QA task. Statistical tests are shown in Appenidx I.

swers $A'_i = \{a'_{i,1}...a'_{i,k}\}$. For the gold answers $A_i = \{a_{i,1}...a_{i,k}\}$, each answer span $a_{i,k} \in P_i$. We follow the input format of UnifiedQA (Khashabi et al., 2020) by concatenating q_i and P_i with a "\n" token. For training labels, we concatenate multiple answers with a ";" token.

6.2 Extractive QA

Given q_i and P_i, this task requires a model to predict whether each token $x_j \in P_i$ is an answer or not. Following the "B-I-O" labeling conventions in the IE field, we create a vector of labels with '2' if x_j is the beginning token of an answer span; '1' if x_j is an internal token of an answer span; '0' if $x_j \notin A_i$. The input is the same as generative QA except that we concatenate q_i and P_i with two "<\s>" tokens to be consistent with the pair-sentence input format of the base model, RoBERTa-large (Liu et al., 2019).

To compare fairly with the generative QA task, we construct candidate answer spans by examining predicted labels for all tokens. Both "BI*" and "I*" cases are considered as valid answers. Finally, we map positive answer tokens' ids back to natural language phrases. More formally, we can denote the final candidate answers of the task as $A''_i = \{a''_{i,1}...a''_{i,k}\}$, where $a''_{i,k} \in P_i$.

6.3 Evaluation Metrics.

It is important to assess how well models can find all valid answer. We evaluate this by using token-based F_1 and exact-match measures. On the other hand, when interacting with machines, we would like the top answer returned to be correct. We measure this by **HIT@1** scores.

- Let U_i, U'_i denotes all uni-grams in A_i, A'_i. We have $F_1^T = \frac{2*P*R}{P+R}$ where $P = \frac{|U_i \cap U'_i|}{|U'_i|}$, $R = \frac{|U_i \cap U'_i|}{|U_i|}$.
- **HIT@1** equals to 1 if the top predicted answer, i.e. $a'_{i,1}$ or $a''_{i,1}$ contains a correct event trigger; otherwise it is 0. This metrics is well defined as all questions in our data contain at least an answer and all (well trained) models return at least one answers. For both generative and extractive QAs, we use the leftmost answer as the top answer.
- **EM** or exact-match equals to 1 if $\forall a'_i \in A'_i, a'_i \in A_i$ **and** $\forall a_i \in A_i, a_i \in A'_i$; otherwise, **EM** = 0.

6.4 Baselines

Model Baselines. For our primary generation QA task, we fine-tuned several sequence-to-sequence pre-trained language models on *ESTER*: BART (Lewis et al., 2020), T5 (Raffel et al., 2020) and UnifiedQA. As mentioned, UnifiedQA (based on BART and T5) is pre-trained on various QA tasks. It also demonstrates powerful zero-shot learning capabilities on unknown QA tasks, which we tested on *ESTER* too. Due to computation constraints, the largest model we are able to finetune is UnifiedQA (T5-large). We leave further investigation to future modeling studies.

Since extractive QA can be considered as a token prediction task, we build our model based on RoBERTa-large with token mask prediction pretraining objectives. Models and fine-tuning details can be found in Appendix G.

Human Baselines. To show the human performance on the task, we randomly select 20 questions

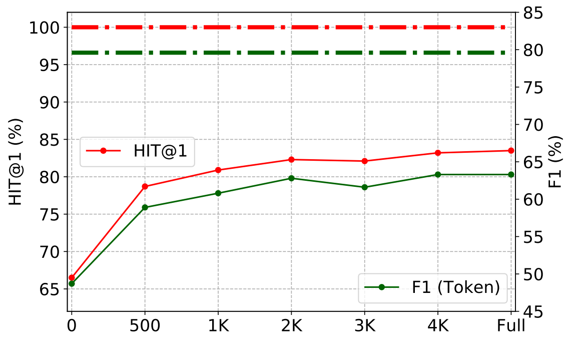


Figure 6: Fine-tuning UnifiedQA-large results by using 500, 1K, 2K, 3K, 4K and full train data. Dashed lines on the top are corresponding human performances.

for each semantic type from the test set. Two co-authors provide answers for these questions, and we compare their mutually agreed answers with the original answers. We ensure co-authors never saw these questions previously. F_1^T, **HIT@1** and **EM** scores are calculated as the human performances.

7 Results and Analysis

In this section, we present and analyze results for the experiments described in Section 6.

7.1 Generative QA

As Table 3 shows, UnifiedQA-large achieves the best average performances among all generative QA baselines, with 63.3%, 83.5% and 22.5% for F_1^T, **HIT@1** and **EM** scores on the test set, which are 16.3%, 16.5% and 13.1% below the human performances. We also observe that UnifiedQA-base with 220M parameters outperforms other comparable or larger models such as T5-base and BART-large with 2-3x more parameters, showing the effectiveness of pre-training with generative QA tasks.

Zero-shot and few-shot Learning. UnifiedQA also demonstrates powerful zero-shot and few-shot learning capabilities in a variety of QA tasks. We observe similar patterns where zero-shot learning from UnifiedQA can significantly outperform its T5 counterpart in Table 3. For few-shot learning, we show in Figure 6 that fine-tuning with only 500-1K examples, the model can achieve quite comparable results with full-training. The model performances level off as the second half of the training data provide $\leq 1.2\%$ improvements across all metrics. This suggests that the benefits of getting more data diminish drastically and data size may not be the bottleneck of learning for *ESTER*.

Breakdown performances. In Figure 7, we show performances for each semantic type on the test data. Not surprisingly, CAUSAL and CONDITIONAL achieve best performances as they are the more dominant semantic types in *ESTER*. Model training may favor these two types. Interestingly, though COUNTERFACTUAL relation has the smallest number of training questions and requires more complex reasoning than CONDITIONAL due to its negation, our models can learn this relation relatively well per **EM** and F_1 measures. This could be contributed by 1) the similarity between COUNTERFACTUAL and CONDITIONAL relations, and 2) the negations are well detected through textual cues in the model training. On the other hand, the significantly lower **HIT@1** score for COUNTERFACTUAL suggests that it is challenging for models to pin-point the most confident answer.

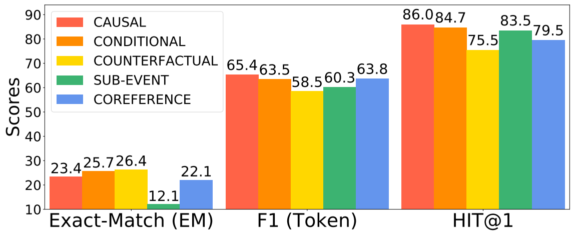


Figure 7: Test performances for each semantic type.

Hierarchical relations, SUB-EVENT and COREFERENCE in general have lower scores than CAUSAL and CONDITIONAL, which could be attributed to two factors: 1) these two categories have smaller percentages (28.1% combined) in training data; 2) understanding these two relations requires complicated reasoning skills to capture not only the hierarchical relations for event triggers but also for their associated arguments. Figure 13 in the appendix shows the similar plateauing effect of adding more training samples for these two relations, which implies that data size may not be the only factor for weaker performances, and these two semantic relations could be inherently challenging to comprehend.

Answer completeness. In Table 2, we show that the validated data contain about 0.5 more answers per question. Besides some rare obvious misses, **proximity** and **saliency** are the two reasons we observe that contribute most to this discrepancy. Our input data include long passages with an average of 128 tokens. Even well-trained workers can overlook relations for event pairs that are physically distant from each other. Moreover, long-distance relations are often less salient. For non-salient relations, expert or external knowledge may be needed to disambiguate. We found workers tend to be conservative by avoiding these non-salient answers.

	#Ans.	F_1^T	**HIT@1**	**EM**
original	1.41	58.7(±0.2)	**78.8(±0.3)**	**18.8(±0.4)**
completed	1.84	**59.3(±0.1)**	78.5(±0.3)	16.9(±0.4)

Table 4: Performances on test data. Workers' original annotations v.s. completed by another worker.

To precisely gauge the impact of answer completeness, we randomly sample 500 questions with type distribution similar to the training data and request qualified workers to find more complete answers. We then retrain UnifiedQA-large with both the original and the more completed answer annotations. Table 4 shows that that the "completed" set has an average number of answers similar to those in our validated data, but we observe no significant improvements. We hypothesize that 1) through our rigorous validation and training, workers are able to identify important answers; 2) the request to find more complete answers could inadvertently introduce some noise, which cancels out the benefits of increasing answer numbers.

7.2 Extractive QA

In this section, we discuss results for the extractive QA task. In Table 3, we observe that extractive QA by finetuning RoBERTa-large achieves the best token F_1 scores, yet under-performs generative QA per **HIT@1** and **EM** metrics. We further compare F_1^T with **EM** scores by increasing training weights on positive tokens, i.e. 'B' or 'I'. Figure 8 shows that as we train models to focus more on the positive answer tokens, F_1^T keeps increasing up to weight = 10, but answer **EM** starts to fall after weight = 2. These results imply that extractive QA excels at finding tokens or phrases that resemble or partially overlap with true answers (good F_1^T scores), but falls short on producing complete and meaningful texts that truly represent event spans.

To verify our hypothesis above, we examine real predictions where both the best generative and extractive QA models do not predict exact answers (i.e. per-sample **EM** = 0). We list several of them in Table 8 of the appendix. In general, extractive QA predicts many single or disconnected tokens that are not meaningful, whereas generative QA, despite making wrong predictions, produces answer spans that are complete and coherent.

To summarize, the comparative studies between generative and extractive QAs emphasize the importance of using multiple metrics to evaluate models and highlight the contribution of leveraging answer generation to solve *ESTER* where complete and meaningful event spans rather than partial tokens are crucial to answer questions.

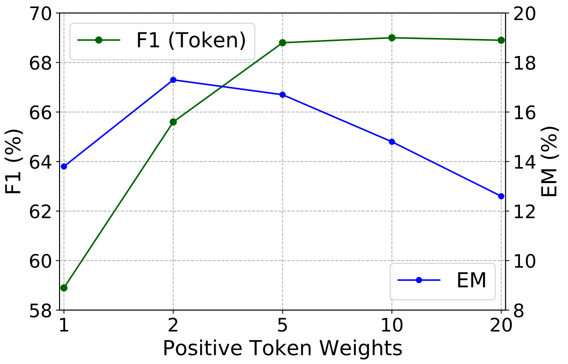


Figure 8: F_1^T v.s. **EM** scores on the dev set by increasing training weights on positive answer tokens.

7.3 Discussion and Future Research

Statistical tests. Modeling is not the main focus of this paper, but we conduct McNemar's tests (McNemar, 1947) for comparable models in Table 7 of Appendix I. Most of the pairwise tests show strong statistical significance.

Future research. *ESTER* facilitates a promising research direction of few-shot learning for event semantic relations as a generative QA task, yet remains challenging since large SOTA systems significantly under-perform human baselines. Future research can explore building question generation systems to automatically annotate a larger scale of data or study the possibilities of transfer learning between this MRC data and other event-centric reasoning tasks.

8 Conclusion

We propose *ESTER*, an MRC datasets for comprehensive event semantic reasoning. We adopt meticulous data quality control to ensure annotation accuracy. *ESTER* enables a generative question answering task, which can be more challenging than the traditional event relation extraction work. The difficulty of the proposed data and task is also manifested by the significant gap between machine and human performances. We thus believe that *ESTER* would be a novel and challenging dataset that empowers future event-centric research.

Acknowledgments

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA), via Contract No. 2019-19051600007 and DARPA under agreement number FA8750-19-2-0500.

References

ACE. 2005. The ace 2005 (ace 05) evaluation plan evaluation of the detection and recognition of ace entities , values , temporal expressions , relations , and events 1.

Gary Carey and Mary Ellen Snodgrass. 1999. *A multicultural dictionary of literary terms*. Jefferson, N.C.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. john wiley & sons.

Goran Glavaš, Jan Šnajder, Marie-Francine Moens, and Parisa Kordjamshidi. 2014. HiEve: A corpus for extracting event hierarchies from news stories. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3678–3683, Reykjavik, Iceland. European Language Resources Association (ELRA).

William Harmon. 2012. *Handbook to Literature*. Pearson.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. *arXiv preprint arXiv:2104.05919*.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

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.

M. O. Lorenz. 1905. Methods of measuring the concentration of wealth. *American Statistical Association*, 9(70.9(70)):209–219.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Pradeep Dasigi, Dheeru Dua, Matt Gardner, Robert L. Logan IV, Ana Marasović, and Zhen Nie. 2020a. Easy, reproducible and quality-controlled data collection with CROWDAQ. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 127–134, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020b. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. *arXiv preprint arXiv:2101.05779*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

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

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Phillip Wolff. 2007. Representing causation. *Journal of Experimental Psychology: General*, 136(1):82–111.

Wenlin Yao, Zeyu Dai, Maitreyi Ramaswamy, Bonan Min, and Ruihong Huang. 2020. Weakly Supervised Subevent Knowledge Acquisition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5345–5356, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Appendix

A Interface

Please refer to Figure 14 for user interface of event selection and Figure 15 for QA annotation.

B QA validity

A pair of QA is valid if and only if it fulfils the following criteria,

1. Both questions and answers MUST contain correct events. Events in questions can have different textual form.

2. Both questions and answers MUST be natural and meaningful. Workers with spotted spamming are immediately disqualified.

3. The semantic relation formed by a QA pair MUST falls into one of the five relation categories we define.

Note that QA validity is different from QA completeness for which we instruct workers to find all possible answers in the passage.

C Type Distribution

Figure 9 & 10 compare the semantic relation type distribution between the train and evaluation data.

IAA Calculation. A pair-wise micro F1 score is calculated by considering one of the annotations as ground truth and the other annotations as predictions. We then rotate the ground truth among all annotations for the same QA and take the average scores as the final IAA scores. For Fleiss's κ score, we follow the same process described in the Fleiss et al. (2013) to evaluate the final IAA.

D Most Frequent N-grams

Enlarged imagines for frequent n-grams in questions can be found in Figure 12a-12e.

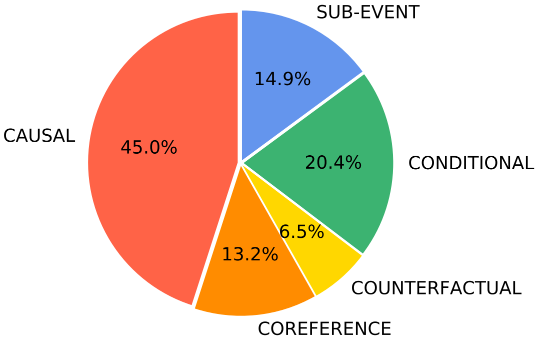


Figure 9: Type Distribution: train data

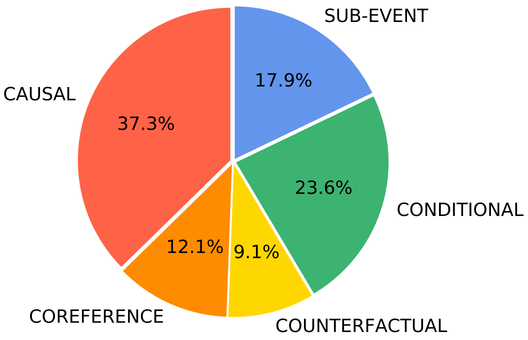


Figure 10: Type Distribution: evaluation data (dev + test)

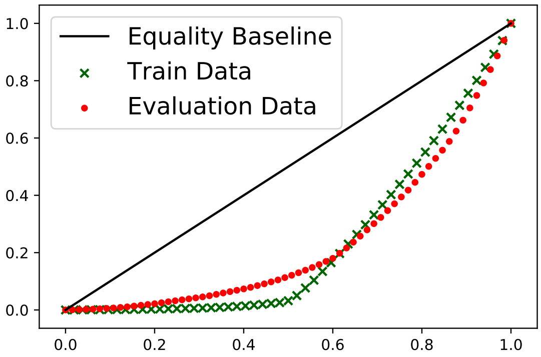
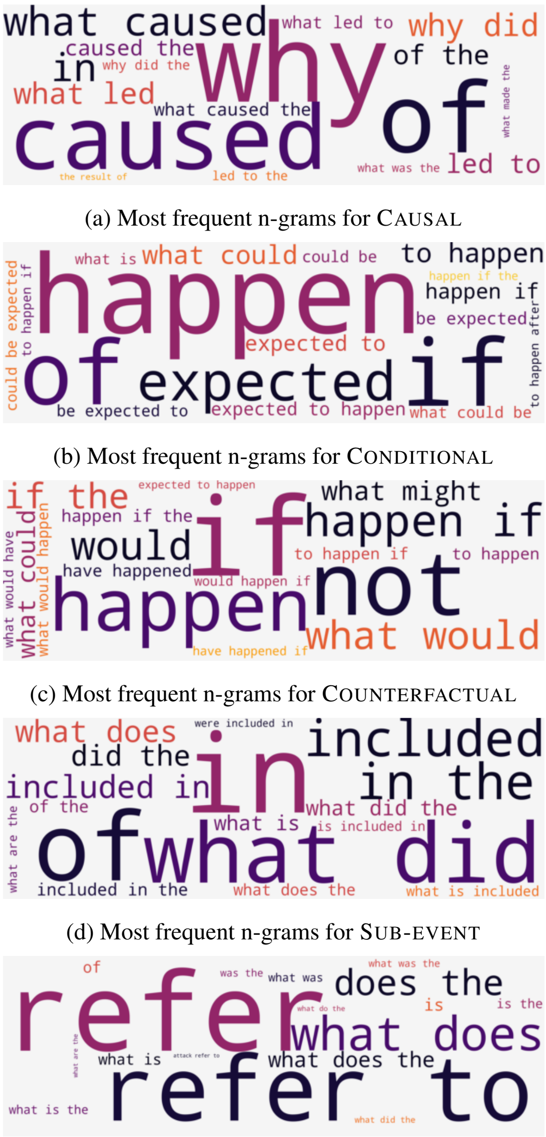


Figure 11: Questions distributions by workers in train v.s. evaluation sets. Equality baseline indicates each participant provides equal number of questions.

E Worker Distribution

We had 70 workers in total who passed our qualification exam and completed at least 1 assignment in our project. Due to our rigorous validating process, only 27 were able to make it into Task 4 and the Large Task which consist of a large number of assignments. Figure 11, known as Lorenze Curve (Lorenz, 1905) illustrates the distribution of number of questions completed by workers. The equality baseline indicates the questions are perfectly well distributed among all workers, i.e. everyone completes the same numbers of questions. The further a curve deviates from the equality baseline, the more unevenly distributed a dataset becomes. Compared with the train data, we observe that the evaluation set is slightly better distributed, which reflects our validation process: for workers who failed our validation tasks and were disqualified, they could still provide some good quality QAs, which we keep in the evaluation data. This increases the diversity of the evaluation set.



(a) Most frequent n-grams for CAUSAL

(b) Most frequent n-grams for CONDITIONAL

(c) Most frequent n-grams for COUNTERFACTUAL

(d) Most frequent n-grams for SUB-EVENT

(e) Most frequent n-grams for CO-REFERENCE

Figure 12: Enlarged charts for most frequent n-grams in questions.

F Number of Tokens.

Table 5 shows an average number of tokens in questions and answers. The COUNTERFACTUAL questions contain the most number of tokens as additional words are often needed to specify the negation reasoning. The average numbers of tokens are all around 6.5 across 5 types of answers. This is exactly the medium of our answer length limits where we set the minimum and maximum numbers of words to be 1 and 12 respectively. The average number of tokens in the passages is 128.1 with the longest passage containing 196 tokens.

G Reproduction Check List

We finetune BART-base, BART-large, T5-base, UnifiedQA-base and UnifiedQA-large on *ESTER*.

	# Tokens	
Semantic Types	Question	Answer
CAUSAL	10.3	6.6
CONDITIONAL	12.1	6.4
COUNTERFACTUAL	13.7	6.0
SUB-EVENT	9.3	6.5
CO-REFERENCE	8.6	6.5

Table 5: Average number of tokens in questions and answers.

UnifiedQA models are all based on T5. Hyper-parameters search ranges are 1) learning rate: $(1e^{-5}, 5e^{-5}, 1e^{-4})$; batch size: $(2, 4)$. Best hyper-parameters can be found in Table 6. We also use 3 random seeds: $(5, 7, 23)$ and report the average performances for each model. For RoBERTa-large, there is an additional hyper-parameter, positive token training weight mentioned in Section 7.2, and it search range is $(1, 2, 5, 10, 20)$.

For BART-base, BART-large, T5-base and UnifiedQA-base models, we were able to finetune on a single Nvidia GTX2020 GPU with 11G memory. For Pegasus and UnifiedQA-large, we have to use a much larger Nvidia A100 GPU with 40G memory. We tried to finetune UnifiedQA based on T5-3B, but we were not able to fit batch size = 1 into a single Nvidia A100 GPU. So we stop at UnifiedQA-large. All reproduction details can be found in the separately submitted code.

Models	# Params.	Best Hyper.	GPU
RoBERTa-large	355M	lr= $1e^{-5}$; b= 2	GTX2080
BART-base	139M	lr= $5e^{-5}$; b= 4	GTX2080
BART-large	406M	lr= $1e^{-5}$; b= 2	GTX2080
T5-base	220M	lr= $1e^{-4}$; b= 4	GTX2080
UnifiedQA-base	220M	lr= $5e^{-5}$; b= 2	GTX2080
UnifiedQA-large	770M	lr= $5e^{-5}$; b= 4	A100

Table 6: Model and fine-tuning details. Learning rate: lr; batch size: b.

H Sub-sample Performances

In Figure 13 we show the fine-tuning UnifiedQA-large using different numbers of training samples for SUB-EVENT and CO-REFERENCE. We observe the same level-off after using 2K training data as in Figure 6 for all semantic types.

I Model Significance Test

To conduct statistical tests over model improvements, we pick the model with highest F_1^T score among the three random seeds for the "best" hyper-

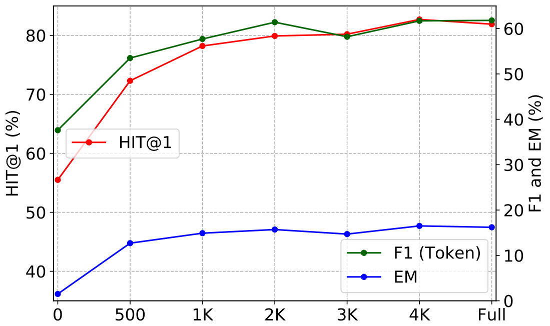


Figure 13: Sub-sample fine-tuning performances for hierarchical relations: SUB-EVENT + CO-REFERENCE. All numbers are average over 3 random seeds.

parameters chosen in Table 6. We then perform McNemar's tests for **HIT@1** and **EM**. Specifically, if **HIT@1** = 1.0 for a sample, we treat it as a correct prediction; otherwise, it is incorrect. The same logic applies to **EM**. We only conduct statistical tests over pairs of models that are comparable in Table 3, and test results are shown in Table 7 below.

Model Comparisons	**HIT@1**	**EM**
Zero-shot		
T5-base → UnifiedQA-base	0.790	≪0.001
UnifiedQA-base → UnifiedQA-large	≪0.001	0.001
Finetune		
BART-base → BART-large	0.865	0.231
T5-base → UnifiedQA-base	0.018	0.574
UnifiedQA-base → UnifiedQA-large	≪0.001	0.022
RoBERTa-large → UnifiedQA-base	≪0.000	0.001
RoBERTa-large → UnifiedQA-large	≪0.000	≪0.000

Table 7: McNemar's test per **HIT@1** and **EM** metrics. Models on the right-hand side of "→" are better. All numbers are p-values with $\leq$ 0.05 indicating statistically significant (underlined).

J Generative v.s. Extractive QA

In Table 8, we show 3 examples comparing predicted answers between generative and extractive QA. In general, scattered answers occur frequently in extractive QA, but barely occur in generative QA. In other words, generative QA is able to consistently produce complete and meaningful answers.

Ex. 1

Passage: the serbs only lifted their threat of a boycott friday after heavy international pressure and the intervention of serbian president slobodan milosevic, a longtime supporter of the rebels.
in a last-minute attempt to get people to vote, the independent democratic serb party (sdss), led by vojislav stanimirovic, launched into what seemed more like a mobilisation rather than a real political campaign.

Question: what could happen if there was no intervention by the serbian president?

Generative Answers: 1. a boycott

Extractive Answers: 1. the; 2. bs; 3. lifted their threat of a boycott

Ex.2

Passage: french defence minister michele alliot-marie on sunday stressed paris's support for the government of lebanese prime minister fuad siniora during a visit to the crisis-wracked nation.
"i have come to reaffirm france's support for the legitimate government of lebanon," she told reporters after meeting her lebanese counterpart elias murr. alliot-marie also stressed paris's backing for the beirut government to "exercise its sovereignty completely", and that the lebanese army play "a role across all its territory". lebanon is undergoing a political crisis with opposition led by shiite movement hezbollah seeking to bring down siniora's government and install a government of national unity. the french minister, who arrived in beirut on saturday for a 48-hour visit, was also to meet siniora before heading to south lebanon for new year's eve with the french contingent of the united nations interim force in lebanon (unifil).

Question: what caused alliot-marie to visit lebanon?

Generative Answers: 1. lebanon is undergoing a political crisis

Extractive Answers: 1. banon is undergoing a political crisis; 2. to; 3. bring down siniora's government

Ex.3

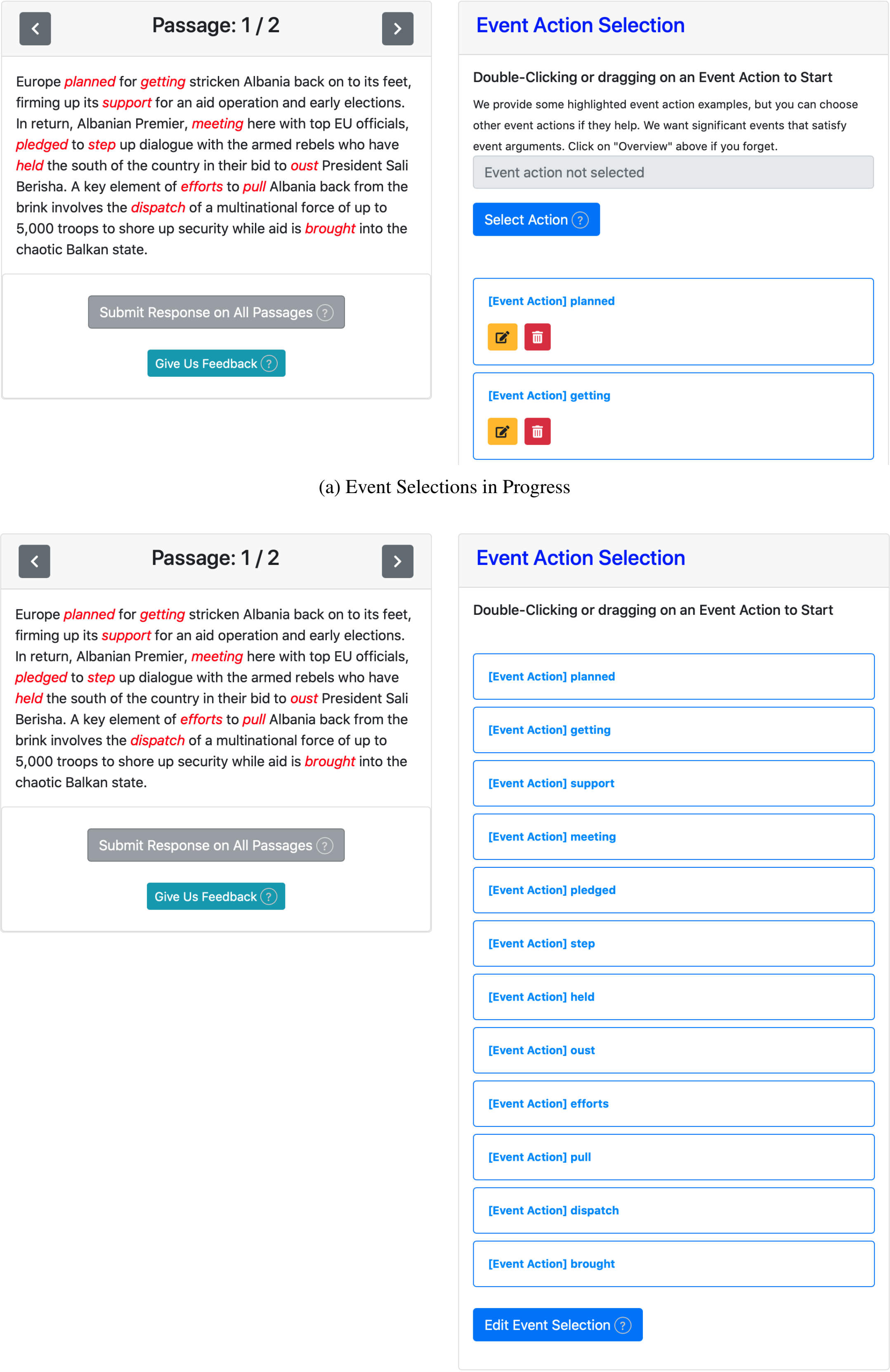
Passage: vieira seems very enthusiastic about bringing in chinese capital and technology into the west african country. he said priorities for bilateral cooperation could expand to ports, roads, bridges and mineral resources. inspired by vieira's enthusiasm, cmec vice president zhou li promised that a special team would fly to guinea-bissau to discuss the details. vieira reminded her that apart from guinea-bissau, other west african countries such as senegal and guinea also need power-generation facilities badly. regarding china as a strategic friend who offers aids without political strings, many african countries impressed with the country's two-digit economic growth are seizing time to explore cooperative opportunities during their stay in beijing to boost domestic economy.

Question: what does the bilateral cooperation include?

Generative Answers: 1. bringing in chinese capital and technology; 2. a special team would fly to guinea-bissau; 3. talk about the details; 4. explore cooperative opportunities during their stay in beijing

Extractive Answers: 1. bringing in chinese capital and technology; 2. expand to; 3. ports, roads, bridges and mineral resources; 4. a; 5. special team would fly to guinea-bissau; 6. discuss the details; 7. explore

Table 8: Examples of answers predicted by generative v.s. extractive QA models. Some passages are shortened for demonstration purpose. Incomplete predictions from extractive QA are highlighted.



Europe *planned* for *getting* stricken Albania back on to its feet, firming up its *support* for an aid operation and early elections. In return, Albanian Premier, *meeting* here with top EU officials, *pledged* to *step* up dialogue with the armed rebels who have *held* the south of the country in their bid to *oust* President Sali Berisha. A key element of *efforts* to *pull* Albania back from the brink involves the *dispatch* of a multinational force of up to 5,000 troops to shore up security while aid is *brought* into the chaotic Balkan state.

Double-Clicking or dragging on an Event Action to Start

We provide some highlighted event action examples, but you can choose other event actions if they help. We want significant events that satisfy event arguments. Click on "Overview" above if you forget.

Event action not selected

[Event Action] planned

[Event Action] getting

(a) Event Selections in Progress

[Event Action] planned

[Event Action] getting

[Event Action] support

[Event Action] meeting

[Event Action] pledged

[Event Action] step

[Event Action] held

[Event Action] oust

[Event Action] efforts

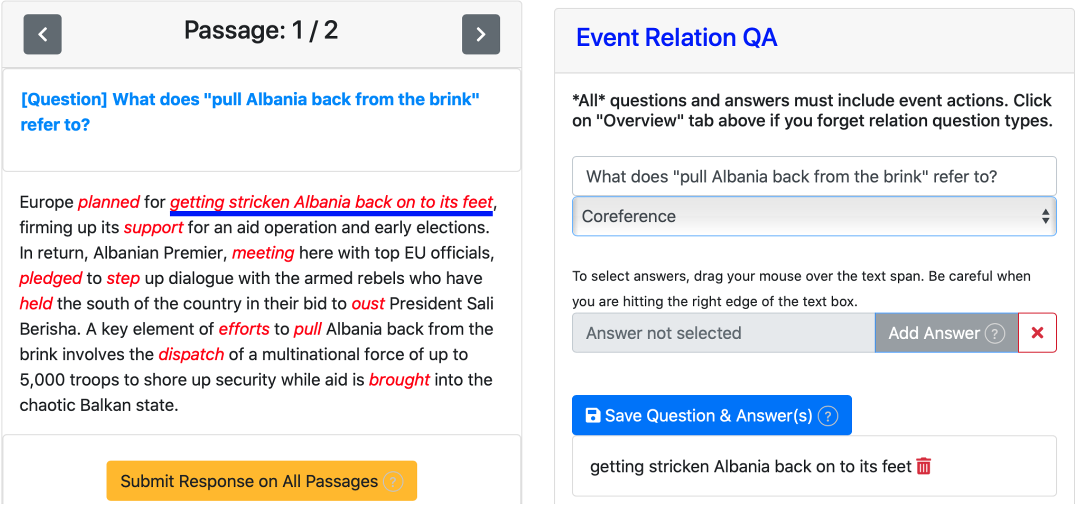
[Event Action] pull

[Event Action] dispatch

[Event Action] brought

(b) Event Selections Completed

Figure 14: An Illustration of Event Selection Interface



Passage: 1 / 2

[Question] What does "pull Albania back from the brink" refer to?

Europe *planned* for *getting stricken Albania back on to its feet*, firming up its *support* for an aid operation and early elections. In return, Albanian Premier, *meeting* here with top EU officials, *pledged* to *step* up dialogue with the armed rebels who have *held* the south of the country in their bid to *oust* President Sali Berisha. A key element of *efforts* to *pull* Albania back from the brink involves the *dispatch* of a multinational force of up to 5,000 troops to shore up security while aid is *brought* into the chaotic Balkan state.

Event Relation QA

All questions and answers must include event actions. Click on "Overview" tab above if you forget relation question types.

What does "pull Albania back from the brink" refer to?

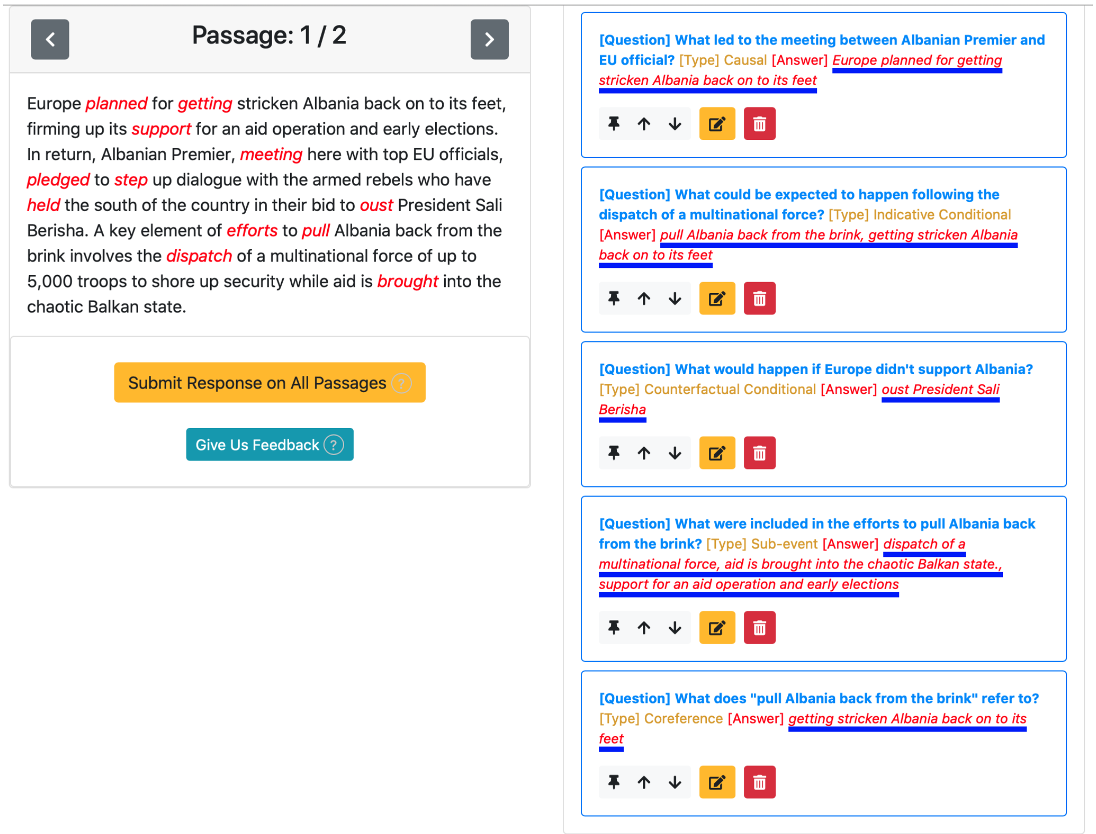
Coreference

To select answers, drag your mouse over the text span. Be careful when you are hitting the right edge of the text box.

Answer not selected

getting stricken Albania back on to its feet

(a) QA Annotations in Progress



Passage: 1 / 2

Europe *planned* for *getting* stricken Albania back on to its feet, firming up its *support* for an aid operation and early elections. In return, Albanian Premier, *meeting* here with top EU officials, *pledged* to *step* up dialogue with the armed rebels who have *held* the south of the country in their bid to *oust* President Sali Berisha. A key element of *efforts* to *pull* Albania back from the brink involves the *dispatch* of a multinational force of up to 5,000 troops to shore up security while aid is *brought* into the chaotic Balkan state.

[Question] What led to the meeting between Albanian Premier and EU official? [Type] Causal [Answer] *Europe planned for getting stricken Albania back on to its feet*

[Question] What could be expected to happen following the dispatch of a multinational force? [Type] Indicative Conditional [Answer] *pull Albania back from the brink, getting stricken Albania back on to its feet*

[Question] What would happen if Europe didn't support Albania? [Type] Counterfactual Conditional [Answer] *oust President Sali Berisha*

[Question] What were included in the efforts to pull Albania back from the brink? [Type] Sub-event [Answer] *dispatch of a multinational force, aid is brought into the chaotic Balkan state., support for an aid operation and early elections*

[Question] What does "pull Albania back from the brink" refer to? [Type] Coreference [Answer] *getting stricken Albania back on to its feet*

(b) QA Annotations Completed

Figure 15: An Illustration of QA Interface