How do we answer complex questions: Discourse structure of long form answers

Anonymous ACL submission

Abstract

Long form answers, consisting of multiple sentences, can provide nuanced and comprehensive answers to a broader set of questions. However, little prior work exists on this task. To better understand this complex task, we study the functional structure of long form answers on two datasets, Natural Questions (Kwiatkowski et al., 2019) and ELI5 (Fan et al., 2019). Our main goal is to understand how humans organize information to craft complex answers. We develop an ontology of sentence-level functional roles for long form answers, and annotate 3.3k sentences in 542 examples. Our annotated data enables training a reliable role classifier that can be used for automatic analysis and thus reveals machine generated answers are structured worse than human written answers. Our data further yields an extractive summarization dataset for long form answers, giving models the ability to identify a concise answer to a complex query.

1 Introduction

While many information seeking questions can be answered by a short text span, requiring a short span answer significantly limits the types of questions that can be addressed as well as the extent of information that can be conveyed. Recent work (Fan et al., 2019; Krishna et al., 2021) explored long form answers, where answers can be free-form texts consisting of multiple sentences. Their multi-sentence nature leads to interesting and nuanced *discourse* within the answers, where the answerer can provide information, hedge, explain, provide examples, point to other sources, and more. Answerers can flexibly structure and organize these elements to provide a coherent, concise answer.

The complexity and flexibility of long form answers pose fresh challenges to the evaluation of long form question answering systems, in stark contrast to short span-based answers where matching spans (Rajpurkar et al., 2016; Joshi et al., 2017) provides a reliable proxy. A recent study (Krishna et al., 2021) demonstrated that automatic metrics like ROUGE (Lin, 2004) are not meaningful for this task and can be easily gamed. Our experiments find that even reliable human preference testing is challenging given the complexity of long form answers, which motivates us to look into the discourse structure of long form answers.

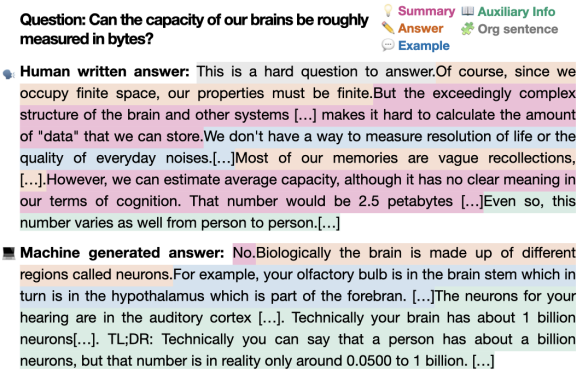


Figure 1: An example of human written and machine generated long-form answer to the same question, annotated with sentence level role. Both are fluent, yet human written answer contains more relevant information and clearer discourse structure, while the machine generated answer contains more auxiliary information.

We take a linguistically informed approach with the dual purpose of *(a)* to better understand the structure of long form answers, and *(b)* to assist the evaluation of long-form QA systems. By characterizing the communicative *functions* of sentences in long form answers (which we call **roles**),[1] e.g., signaling the organization of the answer, directly answering the question, giving an example, providing background information, etc., we analyze human-written, and machine-generated long form answers. Furthermore, our framework combines functional structures with the notion of information

[1] Functional structures have been studied in various other domains (discussed in Sections 3 and 8).

salience by designating a role for sentences that convey the main message of an answer.

We collect annotations on two datasets, ELI5 (Fan et al., 2019) and Natural Questions (NQ) (Kwiatkowski et al., 2019), which contains long form answers written by search users and from Wikipedia page respectively. In total, we provide fine-grained roles for 3.3K sentences (0.5K examples) and coarse annotation for 6K sentences (1.3K examples). We also annotate a small number (94) of machine-generated answers from a state-of-the-art long form question answering system (Krishna et al., 2021) and provide rich analysis about their respective discourse structures. Our analysis demonstrates that studying answer structure can reveal a significant gap between machine-generated answers and human-written answers. We also present a competitive baseline model for automatic role classification, which performs on par with human agreement when trained with our annotated data. Lastly, our dataset yields a novel extractive summarization dataset, providing a benchmark for studying domain transfer in summarization and enabling QA models to provide concise answers to complex queries. We will release all our data and code at `http://anonymous.co`.

2 Revisiting Human Evaluation of Long Form Answers

Recent work (Krishna et al., 2021) dissected the evaluation of long form answers, showing the limitations of lexical matching based automatic metrics. Given the flexibility of long form answers, they suggest human evaluation would be the most appropriate. Initial work (Fan et al., 2019) showed humans could differentiate good answers from bad answers. We further look into the *reliability* of human evaluation in the context of improved model and multiple human written answers – Can humans consistently choose which long form answer is better than the other?

We conduct A/B testing with the long form answers generated from a state-of-the-art LFQA system (Krishna et al., 2021) achieving a high ROUGE score (23.19), and human written answers (H). Their model uses passage retriever (Guu et al., 2020), and generates answers based on the retrieved passage with a routing transformer model (Roy et al., 2021). We look at answers generated from randomly retrieved passages (M-R) and answers generated from the top retrieved passage (M-P). We sample three types of pairs (H, H), (M-P, M-R), (H, M-P), 50 pairs for each type. Given a pair of answers, annotators are asked to choose among four options—*equally good, equally bad, prefer A, prefer B*. The annotators are linguistics undergraduates, trained for our discourse task (Section 4). We collect preferences from three annotators per answer pair and compute inter annotator agreement. The agreement is low across the board, with Fleiss Kappa (Fleiss and Cohen, 1973) value of 0.20 for (H, H), 0.23 for (H, M-P), and 0.26 (M-P, M-R).This shows that when both answers are fluent, it is hard to conduct reliable humans A/B testing.

We empirically verify that evaluating long-form answers is challenging even for humans, as suggested by Krishna et al. (2021). Humans have to evaluate the correctness and sufficiency of the answer, as well as the quality of lengthy machine-generated text (fluency, coherence, etc). This motivates us to study the discourse structure of long form answers, with the focus of evaluating the quality of lengthy generated texts.

3 Defining Answer Discourse Structure

We study the discourse structure of long form answers based on *functional roles* of sentences in the paragraph. Functional structures characterize the communicative role a linguistic unit plays; as such, they vary across genres as the goals of communication also vary. In scientific or technical articles, these roles can be *background, method, findings* (Kircz, 1991; Liddy, 1991; Mizuta et al., 2006), while in news, they can be *main event* or *anecdotes* (Van Dijk, 2013; Choubey et al., 2020).

These structures are related to, though distinct from, coherence discourse structures (Hobbs, 1985). The latter characterizes how each unit (e.g., adjacent clauses or sentences) *relates* to others through semantic relations such as temporal, causal, etc.; such structures can be trees that hierarchically relate adjacent units (Mann and Thompson, 1988) or graphs (Lascarides and Asher, 2008). In contrast, functional roles describe how information is organized to serve the communication goal, in our case, providing the answer. Functional roles will not only inform theoretically-motivate research in long form question answering, but also as we show in Section 5, reflect the quality of answers depending on how humans and models can understand the roles in an answer sentence.

We developed our ontology from long form an-

Role	Why does salt bring out the flavor in most foods?
Org.	Salt does a couple of things that add to the flavor of foods.
Sum.	**First off, it makes things salty.**
Ans.	That sounds simplistic, but salt is one of the 5 basic tastes, so it tends to taste good simply all by itself.
Sum.	**Secondly, salt will lessen sensitivity to other tastes, especially bitter.**
Ans.	This is somewhat true of all the tastes, but adding salt will balance out a dish and not let one taste overpower the others.
Sum.	**Thirdly, it's been shown to increase that aromatic effects of many types of food.**
Ans	A good deal of your "taste" of a food actually comes from the smell of that food (which is why things tend to taste so bland when you nose is congested, like when you have the flu).

Table 1: An example of question answer pair from ELI5 dataset, annotated with sentence-level role.

swers in online community forum (subreddit *Explain Like I'm Five* (ELI5)), hence answers in different domains (e.g., textbooks) can contain roles beyond our ontology. We describe our six sentence-level discourse roles for long form answers here:

Answer-Summary (Sum), Answer (Ans). An answer sentence directly addresses the question. Here we distinguish between the the main content of the answer (henceforth *answer summary*) vs. sentences which explain or elaborate on the summary, shown in Table 1. The summaries play a more salient role than non-summary answer sentences, and can often suffice by themselves as the answer to the question. This is akin to argumentation structure that hierarchically arranges main claims and supporting arguments (Peldszus and Stede, 2013), and news structure that differentiates between main vs. supporting events (Van Dijk, 2013).

Organizational sentences (Org.) Rather than conveying information of the answer, the major role of an organizational sentence is to inform the reader how the answer will be structured. We found two main types of such sentences; the first signals an upcoming set of items of parallel importance:

> **[A]**: **There are a few reasons candidates with "no chance" to win keep running**. 1) They enjoy campaigning[...]

The other type indicates that part of the answer is upcoming amidst an established flow; in the example below, the answerer used a hypophora:

> **[A]**: It might actually be a mosquito bite. I find the odd mosquito in my house in the winter from time to time, and I'm in Canada.[...] **So why does it happen more often when you shower?** It's largely because [...]

Examples (Ex.) Often people provide examples in answers; these are linguistically distinct from other answer sentences in the sense that they are more specific towards a particular entity, concept, or situation. This pattern of language specificity can also be found in example-related discourse relations (Louis and Nenkova, 2011; Li and Nenkova, 2015), or through entity instantiation (MacKinlay and Markert, 2011):

> **[Q]**: What is it about electricity that kills you?
> **[A]**: [...] **For example, static electricity consists of tens of thousands of volts, but basically no amps.** [...]

We found that examples in human answers are often not signaled explicitly, and often contain hypothetical situations:

> **[Q]**: Were major news outlets established with political bias or was it formed over time?
> **[A]**: [...]This is impossible due to the problem of "anchoring." **Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher**[...]

Auxiliary information (Aux.) These sentences provide information that are related to what is discussed in the answer, but not asked in the question. It could be background knowledge that the answerer deemed necessary or helpful, e.g.,

> **[Q]**: Why is it better to use cloning software instead of just copying and pasting the entire drive?
> **[A]**: **When you install an operating system, it sets up what's called a master file table, which [...] are important for the OS to work properly.** [...] Simply copy-pasting files doesn't copy either of these, meaning if you want to back up an OS installation you should clone the disk instead.

or related content that extends the question, e.g.,

> **[Q]**: what is the difference between mandi and kabsa?
> **[A]**: [...] A popular way of preparing meat is called mandi. [...] **Another way of preparing and serving meat for kabsa is mathbi , where seasoned meat is grilled on flat stones that are placed on top of burning embers.**

Notably, the removal of auxiliary information would still leave the answer itself intact.

Miscellaneous (Misc.) We observe various roles that, although less frequent, show up consistently in human answers. These include sentences that acknowledge the limitation of the answer or specify the scope of the answer; describe the original source of the answer; express sentiment about the question or the answer; and refer to other answers in the platform (examples can be found in A.2.1). We group them into a *miscellaneous* role.

Data	Validity	Summary	Role
NQ	263 (1494)	202 (1077)	131 (698)
ELI5	1035 (6575)	834 (5400)	411 (2674)
Total	1298 (8069)	1036 (6477)	542 (3372)

Table 2: Data Statistics. The first number in each cell corresponds to the number of long form answers, and the second number represents the number of sentences.

4 Data and Annotation

4.1 Source Datasets

We use two existing datasets with long form answers: ELI5 (Fan et al., 2019) and Natural Questions (Kwiatkowski et al., 2019). To make annotation task manageable, we filter answers with more than 15 sentences and those with less than 3 sentences, removing about 40% of examples.

ELI5 ELI5 presents QA pairs where question and answers are constructed from the Reddit forum,[2] and has been used as the main dataset for long form QA research (Jernite, 2020; Krishna et al., 2021). In addition to answers in the original datasets, we annotate small amount of machine generated answers from Krishna et al. (2021). We discuss this data in Section 5 separately; analyses in this section do *not* include this data.

Natural Questions (NQ) NQ contains questions from Google search queries, which is then annotated with span-based answers or paragraph level answers from Wikipedia passages. In open retrieval setting, NQ has been exclusively used for its short span based answers (Lee et al., 2019), removing questions with paragraph level answers. We identify NQ contains fair amount of complex queries, and repurpose it to study long form answers for the first time. Many NQ questions can be answered with a short entity (e.g., *how many episodes in season 2 breaking bad?*), but many others questions require paragraph length answer (e.g., *what does the word china mean in chinese?*). This provides complementary answers compared to ELI5 dataset, as answers are not *written specifically* for the questions but harvested from pre-written Wikipedia paragraphs. Thus, this simulates scenarios where machines retrieve paragraphs that can serve as answers instead of generating them.[3]

QA Pair Validity Upon manual inspection, we found in the datasets some long form answers do not address the question, as identified in the ELI5 paper (Fan et al., 2019) which reports 10% of answers to be insufficient. We further remove examples where questions are nonsensical or have presuppositions rejected by the answer; and for simplicity, cases with more than one sub-questions in the question (e.g., *what Tor is, and why everyone praises it as the king of proxies?*). During our annotation, annotators first determine the validity of the QA pair, and proceed to discourse annotation only if they consider the QA pair valid. Details about invalid QA pair identified is in A.2.2.[4]

4.2 Annotation

Annotators We collect annotations via two channels: US-based crowdsource workers on Amazon Mechanical Turk and undergraduate students majoring in linguistics, who are native speakers in English. We aimed to collect all data from crowdsourcing, but making fine-grained role distinction is challenging for untrained annotators. However, they can reliably identify valid QA pairs and sentences serving the *Answer-Summary* role. Thus, we rely on crowdworkers to identify valid QA pairs and summary sentences for each valid answer. Total of 29 crowdworkers worked on our task.

Our six undergraduate students then provided fine-grained role annotations (including summary) for a subset of QA pairs annotated as valid by crowdworkers. We first qualified and then provided training materials to both groups of annotators. The annotation guideline can be found in A.4. We pay crowd workers $0.5 per example, and our undergraduate annotators $13 / hour.

Annotated Data Table 2 presents the data statistics. We collected validity and summary annotations for over 1K answers through crowdsourcing, and fine-grained role annotations for about half of them. As our tasks are complex and somewhat subjective, we collected three way annotations.

We consider a QA pair valid if all annotated it as valid, and invalid if more than two annotated it as invalid. If two annotators considered valid, we collect one additional annotation and consider it valid if and only if the additional annotator marked it as

[2] https://www.reddit.com/r/explainlikeimfive/

[3] We perform additional filtering for NQ question to identify complex questions. Details can be found in A.3

[4] The categories are not mutually exclusive, and we let annotators to pick any of them when an example belongs to multiple categories.

Data	Validity	Summary			Role					
		3	2	1	Answer	Summary	Auxiliary	Example	Org	Misc
NQ	77%	17%	10%	15%	21%	35%	39%	5%	0.4%	0.1%
ELI5	80%	12%	11%	16%	30%	28%	18%	13%	1%	10%

Table 3: Data statistics. The first column represent the number of answers identified as valid (and percentage over all examples in corresponding datasets). The second column set represents the number of sentences identified as the summary sentence. The column count represents the number of annotators who chose that sentence as the summary sentence. The remaining column represents the proportion of each role in respective datasets.

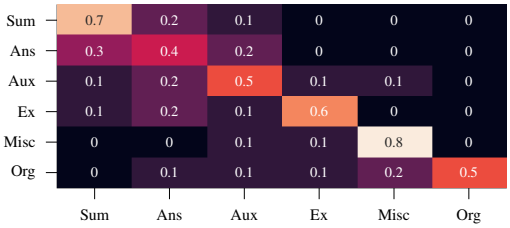


Figure 2: Confusion matrix of role annotations.

valid.[5] We consider the majority role (i.e. chosen by two or more than two annotators) as the gold label. When all annotators chose different roles, they resolved the disagreement through adjudication. We report inter-annotator agreement before the adjudication.

Inter-annotator Agreement We find modest to high agreement for all annotation tasks: For crowdworkers, Fleiss Kappa was 0.53 for summary annotation, 0.51 for validity annotation. For student annotators, Fleiss Kappa was 0.45 for six-way role annotation and 0.52 for summary annotation.

Analysis Table 3 summarizes the label distribution. The proportion of valid QA pairs is similar between the two datasets, yet the role distribution varies significantly. Figure 2 shows the confusion matrix between pairs of annotations, with the numbers normalized by row. We observe frequent confusion between answer vs. answer-summary, and answer vs. auxiliary information.

Around half of the sentences serve roles other than directly answering the questions, such as providing auxiliary information or giving an example. NQ shows higher proportion of auxiliary information, as the paragraphs are written independent of the questions and no miscellaneous sentences.

Figure 3 presents the distribution of each role per its relative location in the answer. Organizational sentences typically locate at the beginning of the answer, examples often in the middle, with an increasing portion of auxiliary information towards the end. Despite the significant differences in the proportion of different discourse roles, the positioning of the roles is similar across NQ and ELI5. We also note a lead bias, as many summary sentences are found at the beginning of the answer.

5 Discourse Structure of Machine Generated Answers

Having observed that humans can reliably assign discourse roles to sentences in (human-written) long form answers, we investigate the discourse structure of machine generated answers. We look into machine generated answers in our initial A/B testing (Section 2) and label them with the same annotation process. These machine generated answers report automatic score (ROUGE-L: 22.74) comparable to that of human written answers we annotate for role (ROUGE-L: 22.2).

We collect validity annotation on 94 machine generated answers, and 42 are considered invalid, among which 40 of them are marked as "no valid answer" by at least one annotator and 29 are marked as so by at least two annotators, suggesting that generated answers can achieve high automatic score *without* answering the question.[6]

We proceed to collect sentence-level role annotations on 52 valid generated long form answers. For the answer role annotation, the annotators *disagree* substantially more as compared to the human-written answers, with a Fleiss kappa of 0.31 (vs. 0.45 for human-written answers), suggesting that the discourse structure of machine generated answers are *less clear*, even to our trained annotators.

The answer role distribution of machine generated answers is very different from that of the human written answers (Figure 4). Machine generated answers contain more sentences which provide auxiliary information, and fewer summary sen-

[5]The validity agreement improves to 0.70 after re-annotation process.

[6]The Fleiss's kappa of QA pair validity is 0.36, substantially lower than the agreement on human written answers (0.51) while annotated by the same set of annotators.

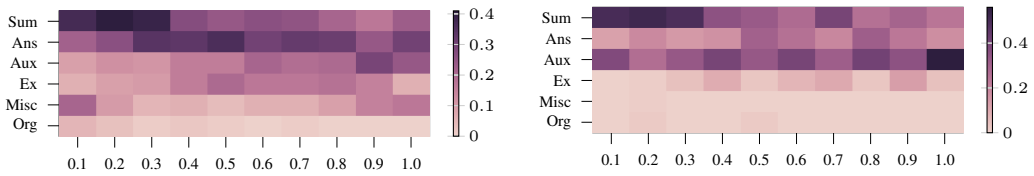


Figure 3: Answer role distribution by the relative position of the sentence in the answer (Left: ELI5, Right: NQ)

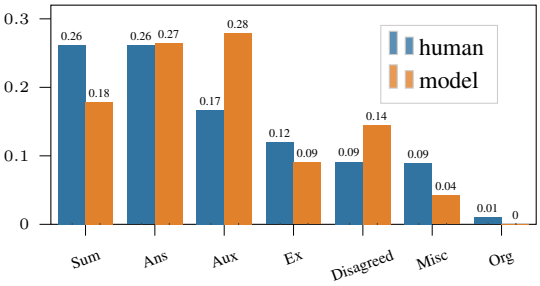


Figure 4: Annotated role distribution of model generated v.s. human written answers for ELI5 dataset, denoted by % sentence.

tences. We also include the portion of "disagreed" sentences where all three annotators chose different roles, which again shows that annotators find the discourse roles of generated sentences confusing. This suggests that model generated answers, despite having high ROUGE scores, is ill-structured.

6 Automatic Discourse Analysis

With the dataset we collected on human written answers, we study how easy it is for models to identify discourse roles for each answer sentence in a valid QA pair.[7] Such a model can enable automatic discourse analysis on long form answers.

Task / Data / Metric Given a question q and its long form answer consisting of sentences $s_1, s_2 ... s_n$, the goal is to assign each answer sentence s_i one of the six roles defined in Section 3.

We randomly split the long form answers in our role annotations into train, validation and test sets with a 70%/15%/15% split. We set apart role annotations for machine generated answers and use those for testing only.

We report accuracy with respect to the majority role label (or adjudicated one, if majority doesn't exist) (Acc), match on any label from any annotator (Match-Any), and Macro-F1 of the six roles.

Lower bounds We present two simple baselines to provide lower bounds: (1) Majority: We predict the most frequent labels in the training data: *Answer-Summary*. (2) Summary-lead: We predict first two sentences as *Answer-Summary*, and the rest of the sentences as *Answer*.

Classification Models We use the [CLS] token from RoBERTa (Liu et al., 2019) which encodes [$question$ <q> ans_1 ... <start> ans_i <end> ...], where ans_i encodes the ith sentence in the answer. The model is trained to predict one of the six roles for ans_i. The model encodes different sentences in the same answer separately.

Seq2Seq We use two variations (base, large) of T5 (Raffel et al., 2019), which take the concatenation of question and answer, and output the roles for each sentence sequentially. We fine-tune the model on our role dataset, by setting the input sequence to be [$question$ [1] ans_1 [2] ans_2 ...], where ans_i denotes the ith sentence in the answer. The target output sequence is set to [[1] $role_1$ [2] $role_2$ [3]...], where $role_i$ is the corresponding role for ans_i.

Human performance We provide two approximations for human performance: upperbound (u) and lowerbound (l). (1) HUMAN (U): We compare each individual annotator's annotation with the majority label. This inflates human performance as one's own judgement affected the majority label. (2) HUMAN (L): We compare pairs of annotation and calculate average F1 and accuracy of all pairs. For Match-any, we compute the match for each annotation against the other two annotations.

Result: Human Written Answers Table 4 reports overall results and Table 5 reports results per each role. T5-large, pre-trained on a large amount of data, shows noticeable gains compared to simple baselines, and is the closest to HUMAN (U). We find Miscellaneous, Example and Summary roles are easier to identify compared to Answer and Auxiliary Information, which are often confused with each other for the annotators as well.

Results: Machine Generated Answers We evaluate our role classifier on the machine generated answers that we annotated, and found that both

[7] We do not automatically classify QA pair validity, as it requires in-depth world knowledge which is beyond the scope of our study.

System	Acc	Match-Any	Macro F1
Majority class	0.30/0.29	0.42/0.44	0.08/0.07
Summary-lead	0.34/0.35	0.57/0.55	0.14/0.15
RoBERTa	0.47/0.48	0.68/0.66	0.46/0.43
T5-base	0.55/0.53	0.74/0.71	0.52/0.54
T5-large	0.57/0.57	0.78/0.76	0.58/0.57
Human (l)	0.58/0.57	0.75/0.74	0.62/0.56
Human (u)	0.77/0.77	1.00	0.79/0.74

Table 4: Role identification result on validation and test answer sentences, presented as (val result/test result).

System	Ans	Sum	Aux	Ex	Org	Msc
RoBERTa	0.38	0.57	0.43	0.56	0.07	0.61
T5-base	0.45	0.56	0.47	0.59	0.42	0.74
T5-large	0.46	0.58	0.52	0.71	0.35	0.80
Human (l)	0.47	0.67	0.54	0.67	0.36	0.77
Human (u)	0.71	0.83	0.75	0.77	0.54	0.87

Table 5: Per role F1 score on the test set.

RoBERTa (Acc: 0.43; Match-Any: 0.60; Macro F1: 0.35) and T5-large model (Acc: 0.50; Match-Any: 0.68; Macro F1: 0.42) report worse performance as compared to human written answers in test set. This echoes our observation that humans agree *less* when annotating machine generated answers.

Usecase: Role Model Uncertainty We further look into whether uncertainty of the model can be used as a proxy to detect poorly organized discourse structure.[8] Our trained role classifier is *less certain* when predicting the roles for machine generated sentences (average entropy on human-written evaluation data / machine-generated answer sentences: 0.81/0.97). Similar to our manual analysis (Section 5), automatic analysis from role classifier sets machine generated answers apart from human written answers.

We plot sentences grouped by their predicted distribution entropy (x-axis) and human agreement (i.e., how many annotators selected the same role during annotation). Figure 5 shows, consistently across human-generated and machine-generated answers, that the more human annotators agree, the lower the classification prediction entropy. This shows that role model entropy can be used to reflect the quality of machine-generated answers without a reference answer, although it wouldn't evaluate the correctness of the answers.

[8]We use RoBERTa for this analysis for simplicity of calculating prediction entropy.

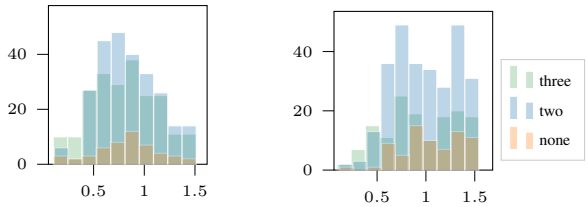


Figure 5: Distribution of role model (RoBERTa) prediction entropy grouped by human agreement on role label (Left: human written answers (test set), Right: machine generated answer. y: # sentence, x: role label entropy).

7 Summarizing Long Form Answers

Lastly, we repurpose our role annotation dataset to summarize long form answers. Finding key sentences from long form answers has a practical appeal as users prefer concise answers (Choi et al., 2021). Our role annotations yield an extractive summarization dataset with 1K answer paragraphs in a new domain of long form answers, with 2.5 sentences marked as summary by one or more annotators.

Task / Data / Metric Given a question q and its long form answer consisting of sentences $s_1, s_2, ...s_n$, the goal is identifying a subset of sentences which can summarize the long form answer.

We merge the answer paragraphs from two datasets and randomly split them into train, validation and test sets with a 70%/15%/15% split.[9] We use all three annotations as gold summary during both training and evaluation, which yields a summarization dataset with multiple references. We therefore report *weighted* precision, recall and F1 scores (Xu et al., 2016). The precise definition of our weighted metric can be found in A.6.

Lower Bounds We present two simple baselines, all taking a fixed number of sentences per paragraph, chosen from [1, 2, 3] based on the performance on validation set. (1) RANDOM: A random set of three answer sentences (2) LEAD: The lead three sentences of the paragraph.

Models We use an extractive summarization model (PreSumm) (Liu and Lapata, 2019) trained on the CNN/DailyMail dataset. PreSumm uses BERT to encode a document and outputs a score for each sentence to determine whether it belongs to the summary or not. We select the threshold of the score based on results on the validation set. We present results on the original model and the

[9]We keep the split aligned with that of the role dataset.

System	P	R	F1
Random	0.36/0.35	0.57/0.58	0.46/0.44
Lead	0.41/0.42	0.65/0.69	0.50/0.52
PreSumm	0.47/0.45	0.61/0.58	0.53/0.51
PreSumm-f	0.47/0.47	**0.80/0.79**	0.59/0.59
T5-sum	**0.61/0.57**	0.78/**0.79**	**0.68/0.66**
Human (a)	0.64*	0.89*	0.74*
Human (m)	0.82*	0.71*	0.76*

Table 6: Summary identification result on validation and test answer paragraphs, presented as (val result/test result). *Human number is computed on a subset.

same model finetuned with our data. We use T5 to identify summary sentences as in role classification, only changing the categories from six-way roles to binary label.

Human performance We approximate human performance with role annotation. Considering 3-way crowdsourced label as gold, we compute performance of summary annotation mapped from the role annotation by undergraduate annotators. We report results from two sets of summary sentences: (1) HUMAN (M): the set of sentence annotated as "Answer-Summary" by more than one annotator. (2) HUMAN (A): the set of sentence annotated as "Answer-Summary" by any annotator.

Result Table 6 reports results on summary task. Lead baseline shows strong performances as was in other domains. The model trained on CNN/Daily mail dataset (Hermann et al., 2015) outperforms lead baseline slightly, but falls behind the model fine-tuned on our dataset. The T5 model fine-tuned on our summary dataset performs the best. The results suggests a significant domain difference between newswire text (where lead is more prominent (Liu and Lapata, 2019)) and long form answers. Thus, our dataset could support future research in extractive summarization across domains.

8 Related Work

Discourse structure. Our work is closely related to functional structures defined through content types explored in other domains; prior work has affirmed the usefulness of these structures in downstream NLP tasks. In news, Choubey et al. (2020) adopted Van Dijk (2013)'s content schema cataloging events (e.g., main event, anecdotal), which they showed to improve the performance of event coreference resolution. In scientific writing, content types (e.g., background, methodology) are shown to be useful for summarization (Teufel and Moens, 2002; Cohan et al., 2018), information extraction (Mizuta et al., 2006; Liakata et al., 2012), and information retrieval (Kircz, 1991; Liddy, 1991). The discourse structure of argumentative texts (e.g., support, rebuttal) (Peldszus and Stede, 2013; Becker et al., 2016; Stab and Gurevych, 2017) has also been applied on argumentation mining. To the best of our knowledge, no prior work has studied the discourse structure of long-form answers.

Question Answering. Recent work (Cao and Wang, 2021) have investigated the ontology of questions, which includes comparison questions, verification questions, judgement questions, etc. We construct the ontology of functional roles of answer sentences. One of the roles in our ontology is summary, yielding an extractive summarization dataset. This shares motivation with a line of work studying query-focused summarization (Xu and Lapata, 2020). Lastly, our work build up on two datasets containing long form answers (Kwiatkowski et al., 2019; Fan et al., 2019) and extends the analysis of long form answers from earlier studies (Krishna et al., 2021).

9 Conclusion

We present a linguistically motivated study of long form answers. We find humans employ various strategies – introducing sentences laying out the structure of the answer, proposing hypothetical and real examples, and summarizing main points – to organize information. Our study also reveals deficient discourse structures of machine-generated answers, showing potential for using discourse analysis to assist in evaluating long form answers in multiple ways. For instance, highlighting summary sentence(s) or sentence-level discourse role could be helpful for human evaluators to dissect long form answers, whose length has been found to be challenging for human evaluation (Krishna et al., 2021). Trained role classifier can also evaluate the discourse structure of machine-generated answers. Future work can explore using sentences belonging to the summary role to design evaluation metrics that focuses on the core parts of the answer (Nenkova and Passonneau, 2004), for assessing the *correctness* of generated the answer. Exploring controllable generation, such as encouraging models to provide summaries or examples, would be another exciting avenue for future work.

References

Maria Becker, Alexis Palmer, and Anette Frank. 2016. Argumentative texts and clause types. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 21–30.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6424–6439, Online. Association for Computational Linguistics.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *ACL*.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint 2002.08909*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Jerry R Hobbs. 1985. On the coherence and structure of discourse.

Yacine Jernite. 2020. Explain anything like i'm five:a model for open domain long form question answering.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint 1705.03551*.

Joost G Kircz. 1991. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. *Journal of documentation*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *NAACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *TACL*.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Elizabeth DuRoss Liddy. 1991. The discourse-level structure of empirical abstracts: An exploratory study. *Information Processing & Management*, 27(1):55–81.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. 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*, abs/1907.11692.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th international joint conference on natural language processing*, pages 605–613.

Andrew MacKinlay and Katja Markert. 2011. Modelling entity instantiations. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 268–274.

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.

Yoko Mizuta, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6):468–487.

Ani Nenkova and Rebecca J Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint 1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Aurko Roy, Mohammad Taghi Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Teun A Van Dijk. 2013. *News as discourse*. Routledge.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yumo Xu and Mirella Lapata. 2020. Query focused multi-document summarization with distant supervision. *ArXiv*, abs/2004.03027.

A Appendix

A.1 Human A/B Testing

We present the agreement for our human evaluation in Table 7, as well as agreement with prior study (Krishna et al., 2021), by calculating agreement for 4-way annotations, including their human evaluation.

A.2 Examples

A.2.1 Miscellaneous Roles

Some sentences specify the limitation of the answer:

> **[Q]**: Why are there such drastic differences in salaries between different countries?
> **[A]**: I'm going to avoid discussing service industries, because[...] **I'm mostly talking tech.** [...]

Some sentences mainly state where the answer came from, e.g.,

> **[Q]**: Why Does a thermostat require the user to switch between heat and cool modes, as opposed to just setting the desired temperature?
> **[A]**: **The person who installed my heat pump (which has all three modes) explained this to me.** [...]

or pointing to other resources:

> **[Q]**: Why did Catholicism embrace celibacy and why did Protestantism reject it?
> **[A]**: **raskhistorians has a few excellent discussions about this.** [...]

Answerers also express sentiment:

> **[Q]**: Why did Catholicism embrace celibacy and why did Protestantism reject it?
> **[A]**: **Good God, the amount of misinformation upvoted is hurting.** [...]

A.2.2 Invalid QA

We provide definitions, as well as examples of each invalid QA types. Table 9 elaborates samples identified as invalid during our annotation.

No valid answer The answer paragraph doesn't provide a valid answer to the question.

> **[Q]**: How does drinking alcohol affect your ability to lose weight?
> **[A]**: Alcohol itself is extremely calorically dense.Doesn't really matter whether you're drinking a light beer or shots, alcohol itself has plenty of calories. Just think of every three shots as eating a mcdouble, with even less nutritional value.

A/B	#	Kappa	Kappa (prior)
Human / Pred	49	0.23	0.24
Random / Pred	49	0.26	0.30
Human / Human	50	0.20	-
Total	148	0.36	-

Table 7: A/B testing results. The second column is the number of answer pairs/

when did the temperance movement begin in the united states
what are the ingredients in chili con carne
is pink rock salt the same as sea salt
why is muharram the first month of the islamic calendar
what qualifies a citizen in the han dynasty to hold a government job
what is the difference between cheddar and american cheese

Table 8: Examples of NQ long questions classified as factoid (top) v.s. non-factoid (bottom).

Nonsensical question The question is nonsensical and it is unclear what is asked.

[Q]: asia vs rest of the world cricket match

Multiple questions asked More than one question are asked in the question sentence.

[Q]: what is a limpet and where does it live

Assumptions in the question rejected The answer focuses on rejecting assumptions in the question, without answering the question.

[Q]: Why is it that as we get older, we are able to handle eating hotter foods
[A]: I'm not sure I accept the premise.Children in cultures where spicy food is common, think nothing of it.My nephews had no problem eating hot peppers when they were very young because it was just a normal part of their diet.[...]

A.2.3 Role annotation

We include example role annotations in Table 10.

A.3 Implementation Details

We use `pytorch-transformers` Wolf et al. (2019) to implement our classification models. The hyperparameters are manually searched by the authors.

Question classification model A difficulty in repurposing NQ is that not all questions with paragraph answers actually need multiple sentences. To identify such complex questions, we built a simple BERT-based classifier, trained to distinguish NQ questions with short answers (i.e., less than five tokens) and ELI5 questions.

Reason	NQ	ELI5
No valid answer	15%	10%
Nonsensical question	1%	0.4%
Multiple questions asked	9%	4%
Assumptions in the question rejected	2%	9%

Table 9: Different reasons for invalid question answer pairs and their frequency in the two datasets.

For the question classification model, we use the [CLS] token from BERT model to perform prediction. We use the original split from the ELI5 dataset, and split the NQ open's validation set into val and test set. We fine-tuned the `bert-base-uncased` model for 3 epochs, with an initial learning rate of $5e-5$ and batch size of 32.We use the model with the highest validation F1 as the question classifier, which achieves F1 of 0.97 and 0.94 on validation and test set respectively. We then run this classifier to select the non factoid questions from NQ questions with long form answers, which classifies around 10%, out of the 27,752 NQ long questions as non-factoid. Examples are in Table 8.

Role classification model For RoBERTa classification model, we use the `roberta-large` model. The training batch size is set to 64, with initial learning rate as $5e-5$. The model is optimized with AdamW optimizer and a linear learning rate schedule. We train the model for 10 epochs and report result of the model with best validation accuracy, averaged across three different random seeds.

For Seq2Seq T5 models, we limit the input/output to be 512/128 tokens. For evaluating the predicted roles, we parse the output string and only take the first k roles into account, k being the number of sentences in the answer paragraph. If the model predicted less than k roles, we pad a dummy role for the remaining sentences. We finetune the model with batch size of 16 and initial learning rate of $1e-4$, with AdamW optimizer and a linear learning rate schedule. We train the model for 30 epoches and report result of the model with best validation accuracy, averaged across three different random seeds.

Summary Identification model For the PreSumm model, we fine-tune by continuing training

(a) Question: What are the benefits of marriage in the U.S.?	Role	Sum
I think one of the biggest ones is that your spouse becomes your legal 'next of kin', meaning you can make medical decisions for them, own their property after they die, etc.	Answer	✓
If you aren't married you are not legally a part of that person's life, so any legal or medical decisions would be up to the parents of that individual.	Answer	
That's why marriage equality was important a few years ago.	Auxiliary	
If someone was with their partner for 15 years and then suddenly dropped dead, their partner had better hope their in-laws liked them or even supported the partnership in the first place.	Example	
If not, the parents could just take the house and all the money (provided the person didn't have a will).	Example	
There are probably other benefits, but I think this is one of the big ones..	Answer	

(b) Question: what is the difference of purple and violet	Role	Sum
Purple is a color intermediate between blue and red .	Answer	
It is similar to violet , but unlike violet , which is a spectral color with its own wavelength on the visible spectrum of light , purple is a composite color made by combining red and blue .	Answer	✓
According to surveys in Europe and the U.S. , purple is the color most often associated with royalty , magic , mystery , and piety .	Auxiliary Information	
When combined with pink , it is associated with eroticism , femininity , and seduction .	Auxiliary Information	

Table 10: Question paired with their paragraph level answer. Each sentence in a paragraph level answer is annotated with its role and whether they consists summary or not. (a) is from ELI5 dataset, (b) is from NQ dataset.

from the checkpoint of `BertSumExt`, following the original set up from the paper.

For the T5 model, we use the same hyper-paramters as the role classification models.

For both setup, we report the results on the model with highest validation macro F1.

A.4 Annotation Interface

Figure 6 is the annotation guideline presented to the annotators (we present Step 1 and Step 3 for crowdworkers, Step 2 and Step 3 for student annotators). We didn't capture the extended example section here due to space.

Figure 7 and 8 are screenshots of the annotation interface.

A.5 Role classification experiment results

A.5.1 Per-role metrics

We report detailed per-role metrics for validation and test set in Table 12.

A.5.2 Experiments on RoBERTa models

We report additional experiments (Table 13 and Table 14) that we have conducted with RoBERTa model, with several variations of the input data. We follow the same experiment set up mentioned in Section A.3.

- **Answer sentence only (Ans)**: This model takes the answer sentence as the input. (i.e. ans_i for the ith sentence in the answer).
- **Question, Answer sentence (Ans-q)**: This model takes the answer sentence with question preppended as input. (i.e. [$question$ <q> ans_i] for the ith sentence in the answer).
- **Answer context (Context)**: This model takes the whole answer paragraph as input, with special tokens indicating the sentence being classified. (i.e. [ans_1 ... <start> ans_i <end> ...] for the ith sentence in the answer).
- **Question, Answer context (Context-q)**: This model takes the whole answer paragraph with question preppended as input. This is the setting we reported in Section 6.

A.6 Summary Identification Evaluation Metric

We report *weighted* precision, recall and F1 scores between system selected summary sentences and ground truth annotation, which consists of (s_i, w_i, y_i), where w_i is the weight and y_i is the label. If none of the annotator select s_i as a summary, y_i is 0 and w_i is 3 (all three annotators agree that this is not a summary). Otherwise, w_i equals to the number of annotators who selected sentence s_i as summary sentence and y_i is 1. Assuming model predicts binary decision on whether s_i belongs to summary or not, denoted as $\hat{y}_i$, we compute weighted TP, FP, TN and FN. Using TP as an example, we calculate it by $TP = \sum_{s_i} w_i$, if $y_i == \hat{y}_i$ and $y_i == 1$. We then use these weighted values to compute precision, recall and F1.

Question: difference between prisoner's dilemma and tragedy of the commons	**Role**
Many real - life dilemmas involve multiple players .	Auxiliary
Although metaphorical , Hardin 's tragedy of the commons may be viewed as an example of a multi-player generalization of the PD : Each villager makes a choice for personal gain or restraint .	Summary
The collective reward for unanimous (or even frequent) defection is very low payoffs (representing the destruction of the " commons ") .	Summary
A commons dilemma most people can relate to is washing the dishes in a shared house .	Example
If not, the parents could just take the house and all the money (provided the person didn't have a will).	Example
By not washing dishes an individual can gain by saving his time , but if that behavior is adopted by every resident the collective cost is no clean plates for anyone .	Example
Question: Why do infants lose their minds when they're tired instead of just falling asleep?	
Answer sentence	**Role**
Think how frustrated you feel when it's the middle of the night and you're nervous and can't get back to sleep.	Example
You're kind of tired, but you can't shut off the anxious thoughts.	Example
Focusing on going to sleep is a skill that has to be learned.	Summary
You can *make* any baby go to sleep, but the trick is to have them "choose" to do it.	Answer
If they aren't taught the skill of going to sleep, they won't know how to do it.	Answer

Table 11: Example role annotation (Top: NQ, Bottom: ELI5)

System	**Answer** P/R/F1	**Summary** P/R/F1	**Auxiliary** P/R/F1	**Example** P/R/F1	**Org** P/R/F1	**Misc** P/R/F1
RoBERTa	0.37/0.34/0.35	0.52/0.58/0.54	0.36/0.42/0.38	0.70/0.60/0.64	0.45/0.15/0.18	0.74/0.61/0.67
T5-base	0.41/0.51/0.45	0.60/0.54/0.57	0.60/0.44/0.51	0.61/0.69/0.65	0.19/0.07/0.11	0.77/0.92/0.84
T5-large	0.40/0.51/0.45	0.62/0.61/0.61	0.64/0.47/0.54	0.76/0.62/0.68	0.52/0.26/0.34	0.79/0.89/0.84
Human (pair)	0.51/0.44/0.47	0.65/0.68/0.66	0.51/0.54/0.52	0.66/0.64/0.65	0.56/0.62/0.58	0.75/0.90/0.81
Human (oracle)	0.69/0.73/0.71	0.85/0.80/0.82	0.75/0.76/0.75	0.82/0.80/0.81	0.82/0.78/0.79	0.87/0.91/0.89
RoBERTa	0.40/0.35/0.38	0.55/0.58/0.57	0.37/0.51/0.43	0.72/0.46/0.56	0.06/0.11/0.07	0.72/0.53/0.61
T5-base	0.40/0.51/0.45	0.58/0.54/0.56	0.52/0.43/0.47	0.61/0.58/0.59	0.35/0.56/0.42	0.80/0.70/0.74
T5-large	0.42/0.50/0.46	0.58/0.58/0.58	0.60/0.47/0.52	0.72/0.71/0.71	0.31/0.56/0.35	0.88/0.73/0.80
Human (pair)	0.47/0.48/0.47	0.67/0.68/0.67	0.51/0.58/0.54	0.64/0.68/0.67	0.36/0.39/0.36	0.80/0.74/0.77
Human (oracle)	0.72/0.70/0.71	0.82/0.84/0.83	0.73/0.77/0.75	0.83/0.74/0.77	0.42/0.78/0.54	0.91/0.84/0.87

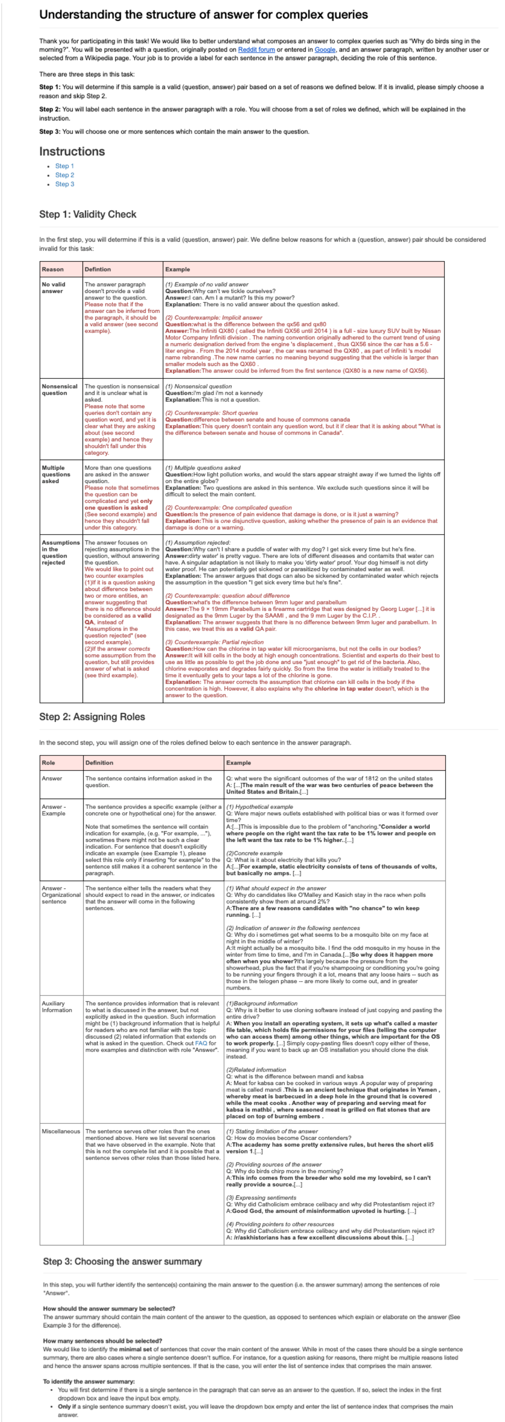
Table 12: Per role performance on validation (top) and test (bottom) set.

System	**Acc**	**Match-Any**	**Macro F1**
Ans	0.45/0.41	0.61/0.61	0.46/0.45
Ans-q	0.46/0.45	0.66/0.63	0.46/0.45
Context	0.47/0.46	0.65/0.64	0.45/0.42
Context-q	0.47/0.48	0.68/0.66	0.46/0.43

Table 13: Role identification result on validation and test answer sentences, presented as (val result/test result).

System	**Ans**	**Sum**	**Aux**	**Ex**	**Org**	**Msc**
Ans	0.24/0.29	0.51/0.42	0.41/0.39	0.42/0.47	0.36/0.38	0.81/0.76
Ans-q	0.30/0.34	0.52/0.45	0.39/0.39	0.47/0.45	0.29/0.32	0.83/0.77
Context	0.31/0.32	0.56/0.55	0.40/0.42	0.57/0.44	0.20/0.11	0.63/0.69
Context-q	0.35/0.38	0.54/0.57	0.38/0.43	0.64/0.56	0.18/0.07	0.67/0.61

Table 14: Per role F1 score on the validation and test answer sentences, presented as (val result / test result).



Understanding the structure of answer for complex queries

Thank you for participating in this task! We would like to better understand what composes an answer to complex queries such as "Why do birds sing in the morning?". You will be presented with a question, originally posted on Reddit forum or entered in Google, and an answer paragraph, written by another user or selected from a Wikipedia page. Your job is to provide a label for each sentence in the answer paragraph, deciding the role of this sentence.

There are three steps in this task:

Step 1: You will determine if this sample is a valid (question, answer) pair based on a set of reasons we defined below. If it is invalid, please simply choose a reason and skip Step 2.

Step 2: You will label each sentence in the answer paragraph with a role. You will choose from a set of roles we defined, which will be explained in the instruction.

Step 3: You will choose one or more sentences which contain the main answer to the question.

Instructions

- Step 1
- Step 2
- Step 3

Step 1: Validity Check

In the first step, you will determine if this is a valid (question, answer) pair. We define below reasons for which a (question, answer) pair should be considered invalid for this task:

Reason	Defintion	Example
No valid answer	The answer paragraph doesn't provide a valid answer to the question. Please note that if the answer can be inferred from the paragraph, it should be a valid answer (see second example).	*(1) Example of no valid answer* **Question:**Why can't we tickle ourselves? **Answer:**I can. Am I a mutant? Is this my power? **Explanation:** There is no valid answer about the question asked. *(2) Counterexample: Implicit answer* **Question:**what is the difference between the qx56 and qx80 **Answer:**The Infiniti QX80 (called the Infiniti QX56 until 2014) is a full - size luxury SUV built by Nissan Motor Company Infiniti division . The naming convention originally adhered to the current trend of using a numeric designation derived from the engine 's displacement , thus QX56 since the car has a 5.6 - liter engine . From the 2014 model year , the car was renamed the QX80 , as part of Infiniti 's model name rebranding .The new name carries no meaning beyond suggesting that the vehicle is larger than smaller models such as the QX60 . **Explanation:**The answer could be inferred from the first sentence (QX80 is a new name of QX56).
Nonsensical question	The question is nonsensical and it is unclear what is asked. Please note that some queries don't contain any question word, and yet it is clear what they are asking about (see second example) and hence they shouldn't fall under this category.	*(1) Nonsensical question* **Question:**i'm glad i'm not a kennedy **Explanation:**This is not a question. *(2) Counterexample: Short queries* **Question:**difference between senate and house of commons canada **Explanation:**This query doesn't contain any question word, but it if clear that it is asking about "What is the difference between senate and house of commons in Canada".
Multiple questions asked	More than one questions are asked in the answer question. Please note that sometimes the question can be complicated and yet **only one question is asked** (See second example) and hence they shouldn't fall under this category.	*(1) Multiple questions asked* **Question:**How light pollution works, and would the stars appear straight away if we turned the lights off on the entire globe? **Explanation:** Two questions are asked in this sentence. We exclude such questions since it will be difficult to select the main content. *(2) Counterexample: One complicated question* **Question:**Is the presence of pain evidence that damage is done, or is it just a warning? **Explanation:**This is *one* disjunctive question, asking whether the presence of pain is an evidence that damage is done or a warning.
Assumptions in the question rejected	The answer focuses on rejecting assumptions in the question, without answering the question. We would like to point out two counter examples (1)If it is a question asking about difference between two or more entities, an answer suggesting that there is no difference should be considered as a **valid QA**, instead of "Assumptions in the question rejected" (see second example). (2)If the answer *corrects* some assumption from the question, but still provides answer of what is asked (see third example).	*(1) Assumption rejected:* **Question:**Why can't I share a puddle of water with my dog? I get sick every time but he's fine. **Answer:**dirty water' is pretty vague. There are lots of different diseases and contamits that water can have. A singular adaptation is not likely to make you 'dirty water' proof. Your dog himself is not dirty water proof. He can potentially get sickened or parasitized by contaminated water as well. **Explanation:** The answer argues that dogs can also be sickened by contaminated water which rejects the assumption in the question "I get sick every time but he's fine". *(2) Counterexample: question about difference* **Question:**what's the difference between 9mm luger and parabellum **Answer:**The 9 × 19mm Parabellum is a firearms cartridge that was designed by Georg Luger [...] it is designated as the 9mm Luger by the SAAMI , and the 9 mm Luger by the C.I.P. . **Explanation:** The answer suggests that there is no difference between 9mm luger and parabellum. In this case, we treat this as a **valid** QA pair. *(3) Counterexample: Partial rejection* **Question:**How can the chlorine in tap water kill microorganisms, but not the cells in our bodies? **Answer:**It will kill cells in the body at high enough concentrations. Scientist and experts do their best to use as little as possible to get the job done and use "just enough" to get rid of the bacteria. Also, chlorine evaporates and degrades fairly quickly. So from the time the water is intitially treated to the time it eventually gets to your taps a lot of the chlorine is gone. **Explanation:** The answer corrects the assumption that chlorine can kill cells in the body if the concentration is high. However, it also explains why the **chlorine in tap water** doesn't, which is the answer to the question.

Step 2: Assigning Roles

In the second step, you will assign one of the roles defined below to each sentence in the answer paragraph.

Role	Definition	Example
Answer	The sentence contains information asked in the question.	Q: what were the significant outcomes of the war of 1812 on the united states A: [...]**The main result of the war was two centuries of peace between the United States and Britain.**[...]
Answer - Example	The sentence provides a specific example (either a concrete one or hypothetical one) for the answer. Note that sometimes the sentence will contain indication for example, (e.g. "For example, ..."), sometimes there might not be such a clear indication. For sentence that doesn't explicitly indicate an example (see Example 1), please select this role only if inserting "for example" to the sentence still makes it a coherent sentence in the paragraph.	*(1) Hypothetical example* Q: Were major news outlets established with political bias or was it formed over time? A:[...]This is impossible due to the problem of "anchoring."**Consider a world where people on the right want the tax rate to be 1% lower and people on the left want the tax rate to be 1% higher.**.[...] *(2)Concrete example* Q: What is it about electricity that kills you? A:[...]**For example, static electricity consists of tens of thousands of volts, but basically no amps.** [...]
Answer - Organizational sentence	The sentence either tells the readers what they should expect to read in the answer, or indicates that the answer will come in the following sentences.	*(1) What should expect in the answer* Q: Why do candidates like O'Malley and Kasich stay in the race when polls consistently show them at around 2%? A:**There are a few reasons candidates with "no chance" to win keep running.** [...] *(2) Indication of answer in the following sentences* Q: Why do i sometimes get what seems to be a mosquito bite on my face at night in the middle of winter? A:It might actually be a mosquito bite. I find the odd mosquito in my house in the winter from time to time, and I'm in Canada.[...]**So why does it happen more often when you shower?**It's largely because the pressure from the showerhead, plus the fact that if you're shampooing or conditioning you're going to be running your fingers through it a lot, means that any loose hairs -- such as those in the telogen phase -- are more likely to come out, and in greater numbers.
Auxiliary Information	The sentence provides information that is relevant to what is discussed in the answer, but not explicitly asked in the question. Such information might be (1) background information that is helpful for readers who are not familiar with the topic discussed (2) related information that extends on what is asked in the question. Check out FAQ for more examples and distinction with role "Answer".	*(1)Background information* Q: Why is it better to use cloning software instead of just copying and pasting the entire drive? A: **When you install an operating system, it sets up what's called a master file table, which holds file permissions for your files (telling the computer who can access them) among other things, which are important for the OS to work properly.** [...] Simply copy-pasting files doesn't copy either of these, meaning if you want to back up an OS installation you should clone the disk instead. *(2)Related information* Q: what is the difference between mandi and kabsa A: Meat for kabsa can be cooked in various ways .A popular way of preparing meat is called mandi .**This is an ancient technique that originates in Yemen , whereby meat is barbecued in a deep hole in the ground that is covered while the meat cooks . Another way of preparing and serving meat for kabsa is mathbi , where seasoned meat is grilled on flat stones that are placed on top of burning embers .**
Miscellaneous	The sentence serves other roles than the ones mentioned above. Here we list several scenarios that we have observed in the example. Note that this is not the complete list and it is possible that a sentence serves other roles than those listed here.	*(1) Stating limitation of the answer* Q: How do movies become Oscar contenders? A:**The academy has some pretty extensive rules, but heres the short eli5 version 1**.[...] *(2) Providing sources of the answer* Q: Why do birds chirp more in the morning? A:**This info comes from the breeder who sold me my lovebird, so I can't really provide a source.**[...] *(3) Expressing sentiments* Q: Why did Catholicism embrace celibacy and why did Protestantism reject it? A:**Good God, the amount of misinformation upvoted is hurting.** [...] *(4) Providing pointers to other resources* Q: Why did Catholicism embrace celibacy and why did Protestantism reject it? A: **/r/askhistorians has a few excellent discussions about this.** [...]

Step 3: Choosing the answer summary

In this step, you will further identify the sentence(s) containing the main answer to the question (i.e. the answer summary) among the sentences of role "Answer".

How should the answer summary be selected?
The answer summary should contain the main content of the answer to the question, as opposed to sentences which explain or elaborate on the answer (See Example 3 for the difference).

How many sentences should be selected?
We would like to identify the **minimal set** of sentences that cover the main content of the answer. While in most of the cases there should be a single sentence summary, there are also cases where a single sentence doesn't suffice. For instance, for a question asking for reasons, there might be multiple reasons listed and hence the answer spans across multiple sentences. If that is the case, you will enter the list of sentence index that comprises the main answer.

To identify the answer summary:

- You will first determine if there is a single sentence in the paragraph that can serve as an answer to the question. If so, select the index in the first dropdown box and leave the input box empty.
- **Only if** a single sentence summary doesn't exist, you will leave the dropdown box empty and enter the list of sentence index that comprises the main answer.

Figure 6: Screenshot of annotation guideline.

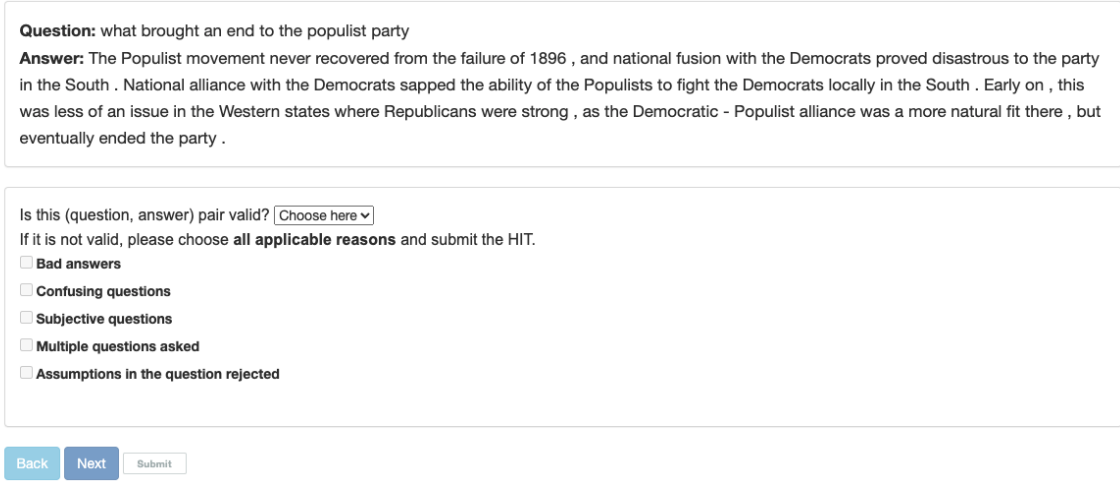


Figure 7: Screenshot of annotation interface for question validity.

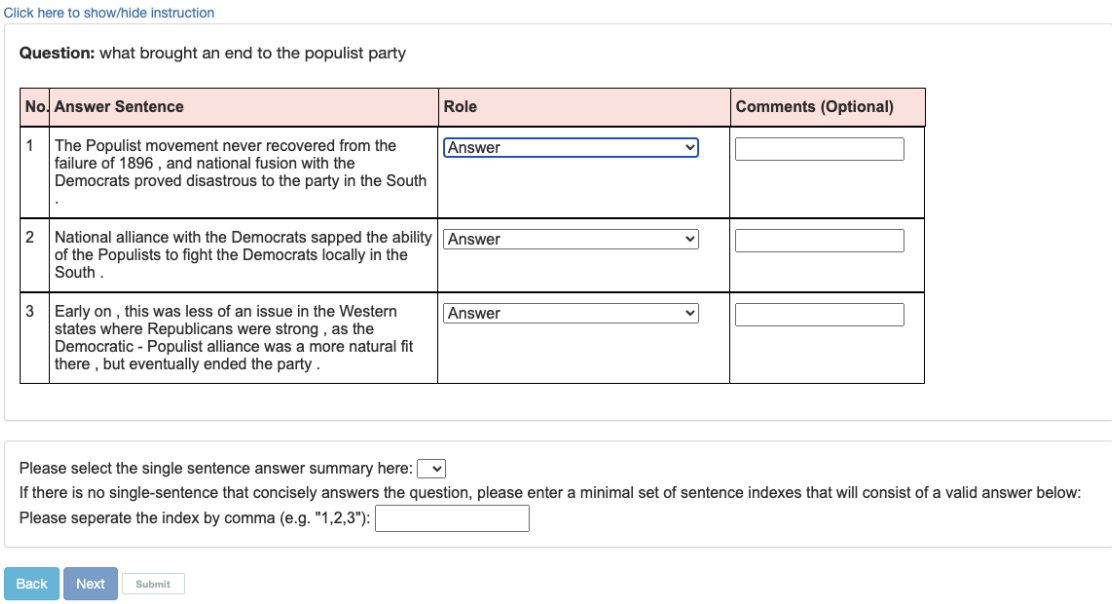


Figure 8: Screenshot of annotation interface for sentence-level role, as well as summary sentence selection.