What are the implications of your question? Non-Information Seeking Question-Type Identification in CNN Transcripts

Anastasiia Tatlubaeva*, Yao Sun*, Zhihan Li*, Chester Palen-Michel
Brandeis University
Waltham, MA
{atatlubaeva, yaosun, zhihanli, cpalenmichel}@brandeis.edu
* These authors contributed equally to this work

Abstract

Non-information seeking questions (NISQ) capture the subtle dynamics of human discourse. In this work, we utilize a dataset of over 1,500 information-seeking question(ISQ) and NISQ to evaluate human and machine performance on classifying fine-grained NISQ types. We introduce the first publicly available corpus focused on annotating both ISQs and NISQs as an initial benchmark. Additionally, we establish competitive baselines by assessing diverse systems, including Generative Pre-Trained Transformer Language models, on a new question classification task. Our results demonstrate the inherent complexity of making nuanced NISQ distinctions. The dataset is publicly available at https://github.com/YaoSun0422/NISQ_dataset.git.

Keywords: non-information seeking questions, question answering, conversational AI, information-seeking queries, annotating, information-seeking questions, OpenAI models, question classification task

1. Introduction

In recent years, Natural Language Processing (NLP) has achieved remarkable progress in recognizing and addressing information-seeking questions (ISQs)—queries where users primarily hunt for specific answers. Yet, a distinct category of questions, termed non-information-seeking questions (NISQs), continues to pose a significant challenge for conversational AI, due to their varied nature and heavy reliance on context. Although they may not directly seek information, NISQs are pivotal in capturing the nuanced dynamics of human discourse. Distinguishing between the various types of NISQs is crucial for refining conversational AI's responsiveness and understanding the subtleties of human communication, leading to more natural interactions and a heightened user experience. Our contributions in this paper are three-fold. First, we introduce the first publicly available corpus focused on annotating non-information-seeking questions, serving as an initial benchmark. Second, we evaluate human and machine performance on classifying fine-grained types of non-information-seeking questions, establishing competitive baselines with models. Third, our results demonstrate the inherent complexity of making nuanced distinctions between non-information-seeking question types.

2. Related work

In the domain of NLP, much emphasis has been placed on understanding and responding to ISQs. There are a number of question answering datasets with ISQs such as SQuAD (Rajpurkar et al., 2016) and CoQA (Reddy et al., 2019) among others. Conversely, NISQs have not been as thoroughly investigated. Our research is informed by the work of Kalouli et al. (2021), which delves into a more comprehensive Question Type Identification (QTI) task, and Kalouli et al. (2018), which offers an insightful review of the various types of NISQs. Although Kalouli et al. (2021) and Kalouli et al. (2018) do discuss the distinction between different types of NISQs, they were pursuing a more general goal of creating a corpus of questions where each question was labeled either as an ISQ or an NISQ. It is worth noting that existing research has covered specific types of NISQs questions individually; for instance, deliberative questions are examined in Wheatley (1955), tag questions in Kim and Ann (2008), and rhetorical questions in Bhattasali et al. (2015), and so on, however, a comprehensive and finer-grained classification of diverse NISQs remains an open area of research.

3. Corpus Creation

Our data came from the RQueT (Resource of Question Types) dataset described in Kalouli et al. (2021). It is a dataset of CNN transcripts of live discussion and interviews from 2006–2015. Each question was labeled as ISQ or NISQ. However, to avoid potential biases in our analysis and maintain objectivity, we chose not to consider the original labels and approached our research with a fresh perspective. Due to copyright reasons, the data can

LREC-COLING 2024, pages 17444–17448

be requested[1] for downloading, and then a script from Kalouli et al. (2021) on their Github repository can be used for compiling the dataset [2]. We will also release our dataset in this manner.

Using the 'profanity-check' Python package, we eliminated entries with inappropriate language and manually corrected those that were improperly parsed. After pre-processing and filtering, our dataset consisted of 1,566 texts.

For each text in the dataset, it has one target question, two sentences before and after the target question, and speaker information. Table 1 shows an excerpt of our corpus.

3.1. Annotation Guidelines

Our initial labels were sourced from the original NISQ types detailed in Kalouli et al. (2021). However, after the initial labeling, we refined and consolidated some of the categories, resulting in a concise set of seven labels.

Here is the definition for each label:

DELIBERATIVE: These questions encourage participants in the conversation to share their perspectives on the topic broached in the question.

RHETORICAL: Typically, rhetorical questions don't seek an answer or already imply one. Their primary function is to underscore the speaker's viewpoint.

ECHO: Used for clarity or to express emotions, these questions often repeat or mirror prior statements, showcasing surprise, confusion, or a request for validation.

TAG: Appended to main statements, they seek affirmation or challenge a premise. These can be straightforward or occasionally sarcastic.

QUOTED: Representing another's inquiry, they capture questions posed by someone other than the current speaker.

OTHER: A category for questions that don't fit the criteria of the above NISQs labels.

ISQ : Designed to obtain factual details from participants.

3.2. Annotation Procedure

Four graduate students specializing in computational linguistics annotated each question in our corpus. Presented with a text, each annotator was tasked with selecting one label from the seven labels that most accurately represented the nature of the target question. To aid in their decision-making, the annotators were given the definition of each type of question, some examples, and descriptions of each tag in more details, ambiguous cases, and no further instructions, ensuring an objective annotation process. With four annotators, we divided our data into four batches with 391-392 data points and each person annotated two batches (783 data points per person).

3.3. Inter-Annotator Agreement

As we delved into the data, we became aware of the need to adapt our guidelines for more accurate representation, which led us to retain primary labels such as DELIBERATIVE and RHETORICAL. We then combined the less frequent labels into a unified super-label termed 'OTHER'. Consequently, our revised labeling structure encompassed four categories: DELIBERATIVE, RHETORICAL, ISQ, and OTHER.

The final set of labels, as also reflected in the completed corpus, included the types: DELIBERATIVE, RHETORICAL, ISQ, and OTHER.

We computed inter-annotator agreement metrics before merging OTHER to include ECHO, TAG, and QUOTED and after the merger. Agreement metrics including Cohen's Kappa (Cohen, 1960), Bennet's S (Bennett et al., 1954), and Krippendorf's alpha (Krippendorff, 2011) is shown in Table 2.

3.4. Ambiguous example

Labeling question types is intrinsically sensitive to context. For accurate annotation, the process demands substantial information. However, the RQueT (Resource of Question Types) dataset provides only five sentences per conversation, which often proves insufficient for annotators to confidently determine question types. Consequently, annotators rely on their individual interpretations, leading to a plethora of ambiguous cases.

Consider the example in Table 1. In this dialogue, the question posed by Jill Zuckman has multiple interpretations. It may be viewed as a direct inquiry into others' perspectives, categorizing it as DELIBERATIVE. Alternatively, it could be perceived as an exploratory probe into their opinions, aligning it with RHETORICAL. Such ambiguities recurrently emerge during annotation, leading to lower consensus among annotators.

3.5. Adjudication

A team of three adjudicators who wrote and were extensively familiar with the guidelines conducted adjudication. Adjudicators addressed discrepancies in the annotations by first comparing conflicting data against preliminary annotation that had

[1] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ISDPJU

[2] https://github.com/kkalouli/RQueT

Sentence	Text	Speaker
Ctx 2 Before	I mean, voters feel a sense of empathy for what shes been through.	JILL ZUCKMAN
Ctx 1 Before	And so I think that has a lot to do with it.	JILL ZUCKMAN
Question	But at the same time, I think there are questions that are very delicately being raised about, well, if you dont know how much time you have left, and you have these young children, why are you going forward with this?	JILL ZUCKMAN
Ctx 1 After	And shes been – shes been very candid about it, that she doesnt want to let the cancer beat her.	JILL ZUCKMAN
Ctx 2 After	Do you think these questions should be raised delicately or otherwise by the media?	KURTZ, HOST

Table 1: Sample of the corpus format. Each row contains a sentence and its context before and after. The question and its context also hold the speaker information.

Metric	IAA Before	IAA After
Cohen's K	0.10	0.13
Bennet's S	0.27	0.22
Krippendorff's α	0.07	0.10

Table 2: Inter-Annotator Agreement (IAA) before and after the introduction of a superlabel.

Label	Questions
DELIBERATIVE	793
RHETORICAL	341
OTHER	201
ISQ	173
Total	1,508

Table 3: Label distribution in our data.

been conducted previously by members of the adjudication team. If the adjudicator's choice aligned with one of the annotator's selections, that label was adopted. When an adjudicator's label differed from both annotators, we gave precedence to the adjudicator's label. All remaining conflicts without existing adjudicator judgments were then grouped into three batches (281, 279, and 279 questions) for further review. To expedite the process, the final decision was narrowed down to selecting between the two labels originally chosen by annotators.

3.6. Final Dataset

After adjudication and post-processing, we remove the meaningless and repeated conversation and conversation that contains discriminatory words. We ended up with 1,508 data points. Please see Table 3 for a breakdown on data frequency by class. The data was split into training, validation, and test data sets in an 80/10/10 manner.

4. Experiments

We did multi-class classification with four labels: DELIBERATIVE, RHETORICAL, OTHER, and ISQ. For metrics, we considered accuracy and Matthew's correlation coefficient (mcc). The baseline accuracy, if the model always picked the most frequent label (DELIBERATIVE), would be 0.52. The baseline mcc, if the model made random predictions, would be 0.0.

We fine-tuned with two base models RoBERTa (Liu et al., 2019) and BigBird (Zaheer et al., 2020). We experimented with differing amounts of context. The first model was given only the target question, the second one was also given one sentence before and after the target question, and the third one also received two sentences before and after the target question. Since the average input data point for model 2 and 3 was longer than 512 tokens, we used the BigBird model for those models and a regular RoBERTa model for model 1. Our hypothesis was that the more context the model gets from the data, the better its performance will be. For the RoBERTa model we used a learning rate of 1e-4 and a batch size of 128, while for BigBird we used a learning rate of 1e-5 and a batch size of 32. Ten training epochs were used in all settings. Those hyperparameters were selected based on multiple trials and showed the best results.

4.1. Baseline Results

Table 4 shows the results of evaluating all three models on the test set. Our best performing model was model 1. Despite having access only to the target question (which sometimes consisted only of one word), it was able to correctly identify the question type 64% of the time. The performance of models 2 and 3 did not differ significantly, with model 3 showing only 3% improvement from model 2. Both models 2 and 3 performed worse than model 1 (by 13% and 10% respectively), with model 2 not achieving the baseline accuracy. At the same time, the mcc was around 0.3 for mod-

Model	Accuracy	mcc
RoBERTa (Target Only)	0.64	0.38
BigBird (Window Size 1)	0.51	0.31
BigBird (Window Size 2)	0.54	0.32

Table 4: Accuracy and mcc on the test set for all three models.

els 2 and 3 and just below 0.4 for model 1 which shows that the model is far from making perfect predictions (an mcc of 1.0) but also not making random predictions (an mcc of 0.0).

As can be seen from Table 4 with the evaluation results on the test set, our experiment did not confirm our hypothesis and the best performing model was the model that was given the least amount of context. This result is surprising since, based on our experience annotating the data as well as the experience of our annotators, the context played a crucial role in identifying the type of the question.

We believe that the results we got are due to the fact that two utterances before and after the target question are not enough context. In the original paper, Kalouli et al. (2021) point out that one of the most useful features in their experiments was speaker information. In particular, all models they experimented with were able to predict the question type correctly 77% of the time if they were given just one piece of information: the speaker of the target question and the speaker of the first utterance after the target question Kalouli et al. (2021). Although Kalouli et al. (2021) were doing binary classification between ISQs and NISQs, it is still notable that their models did not need to consider the target question itself to achieve high performance. Our models, however, completely ignore speaker information which might have resulted in their relatively poor performance.

4.2. Generative Pre-Trained Transformer Language Model Performance

Because annotation requiring linguistic knowledge is costly, we attempt to determine whether gpt-3.5-turbo, a generative pre-trained transformer large language model provided by OpenAI, can effectively manage subjective question types.

In this experiment, we carefully selected 100 questions that were agreed upon by our three authors and had minimal annotator disagreement for testing. We limited options to the four labels in our final dataset.

We designed a prompt as follows: The first section introduces our task, including the format of our data, the labels employed, and their significance. The second delves into a detailed description of each label, supported by examples. In the third section, we present our task and formulate the question [3].

In our experiment, out of the 100 instances selected, gpt-3.5-turbo accurately identified 42% of the correct labels within 2 minutes. In contrast, when using the same 100 instances with the regular RoBERTa model we mentioned above, the accuracy increased to 59%. This demonstrates that using an LLM may be useful as a pre-annotation step, and that training a smaller task specific model can still outperform LLMs for tasks like QTI that may require finer understanding of nuance and context.

5. Future Work

Our Inter-Annotator Agreement (IAA) indicates the task of identifying NISQs is challenging and requires linguistic expertise and occasional intuition. We found ourselves frequently engaged in extensive discussions over single questions from our dataset, striving to pinpoint their type. Feedback from annotators also echoed similar challenges. Providing longer context to annotators may help with the annotation task, but risks making the task more time consuming for annotators.

As it is sourced exclusively from CNN transcripts, our data is heavily skewed towards political discussions. This bias not only introduced challenges such as potential misinterpretation due to nuanced political rhetoric but also necessitated an in-depth understanding of both US and global political landscapes. To mitigate these challenges and aim for a more generalized understanding, utilizing a dataset that delves into everyday, relatable conversations on broader topics would be a promising future direction for the project.

As highlighted in Section 4.1, our model training currently omits speaker details. Integrating this information could offer a deeper understanding of context, especially when differentiating between speakers' styles or biases. An added layer of context might provide a richer foundation for analysis.

6. Conclusion

In summary, this study marks the debut of the first publicly accessible corpus dedicated to the annotation of both information-seeking and non-information-seeking questions, thereby establishing a foundational benchmark for future research in this domain. Our findings reveal the intricate challenges involved in differentiating between various types of non-information-seeking questions. Looking ahead, there are avenues for enhancing the annotation guidelines, diversifying the dataset,

[3] A full example will be included in an appendix due to length.

and investigating further machine learning methodologies.

7. Bibliographical References

Edward M Bennett, Renee Alpert, and AC Goldstein. 1954. Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Shohini Bhattasali, Jeremy Cytryn, Elana Feldman, and Joonsuk Park. 2015. Automatic identification of rhetorical questions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 743–749.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jong-Bok Kim and Ji-Young Ann. 2008. English tag questions: Corpus findings and theoretical implications. *English Language and Linguistics*, 25:103–126.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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*.

J. M. O. Wheatley. 1955. Deliberative Questions. *Analysis*, 15(3):49–60.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

8. Language Resource References

Kalouli, Aikaterini-Lida and Kaiser, Katharina and Hautli-Janisz, Annette and Kaiser, Georg A. and Butt, Miriam. 2018. *A Multilingual Approach to Question Classification*. European Language Resources Association (ELRA).

Kalouli, Aikaterini-Lida and Kehlbeck, Rebecca and Sevastjanova, Rita and Deussen, Oliver and Keim, Daniel and Butt, Miriam. 2021. *Is that really a question? Going beyond factoid questions in NLP*. Association for Computational Linguistics.

Rajpurkar, Pranav and Zhang, Jian and Lopyrev, Konstantin and Liang, Percy. 2016. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

A. ChatGPT Prompt

Following the official tutorial and after testing several variations, we finalized our prompt as follows:

> "Your task is to label questions. The third sentence of the provided text is a question. Please label it; we've supplied the context before and after for clarity. You will have a choice between the following four labels: Deliberative, Rhetorical, ISQ, and Others.
>
> Here is the description of the label 'Deliberative': Deliberative questions invite other people in the conversation to contribute their ideas on the topic expressed in the question. Here is an example:
>
> (Ctx 2 Before) HOWARD KURTZ: Back in July, when Terry Jones had tweeted he was going do this and it got a little of attention, most Americans were aware of this – we saw the clip at the top of the show, Rick Sanchez putting him on CNN. (Ctx 1 Before) HOWARD KURTZ: Should he have done that? (Question)HOWARD KURTZ: Why does Terry Jones warrant any air time at all? (Ctx 1 After) DAVID FRUM: Well, it is exciting, and that is a kind of tabloidy show. (Ctx 2 After) DAVID FRUM: And you hope – theres a part I think of every journalists mind that sort of hopes for a big global reaction.
>
> (Due to space constraints, we have not shown all the descriptions and examples for all the four labels)
>
> The text is: (given the text here). What is the label of the text?"