Narrative Why-Question Answering: A Review of Challenges and Datasets

Emil Kalbaliyev
Institute of Computer Science
University of Tartu
Tartu, Estonia
emil.kalbaliyev@ut.ee

Kairit Sirts
Institute of Computer Science
University of Tartu
Tartu, Estonia
kairit.sirts@ut.ee

Abstract

Narrative Why-Question Answering is an important task to assess the causal reasoning ability of models in narrative settings. Further progress in this domain requires clear identification of challenges that question answering models need to address. Since Narrative Why-Question Answering combines the characteristics of both narrative understanding and why-question answering, we review the challenges related to these two domains. In the context of why-questions, we review the characteristics of causal relations and the sources of ambiguity in why-questions. In relation to narratives, we discuss the challenges posed by the implicitness and the length of the narrative texts. Furthermore, we identify suitable datasets for Narrative Why-Question Answering and outline both data-specific and task-specific challenges that can be utilized to test the performance of models. Additionally, we discuss some issues that can pose problems in benchmarking Narrative Why-Question Answering systems.

1 Introduction

Narrative Why-Question Answering is the task of answering why-questions in narrative settings. This task combines the challenging properties of both why-question answering and narrative understanding. As such, Narrative Why-Question Answering makes a suitable task for evaluating complex comprehension abilities of language models.

Why-question is one of the most challenging non-factoid question types (Bolotova et al., 2022) because it requires discovering explicitly or implicitly stated causal relations from text. As such, why-questions can be used to test the causal reasoning abilities of QA systems. On the other hand, narratives can be considered a desirable testbed for machine reading comprehension (MRC) tasks because narratives play a central role in the life of human beings, they have implicit nature and complex structure, and fictional narratives are self-contained (Dunietz et al., 2020).

Although humans can easily identify causal relations in narratives and make inferences by using their background knowledge and by paying close attention to the timeline, cause, and motivation of the events/entities, current QA systems have difficulties extracting correct relations and making such complex inferences in narratives (Lal et al., 2021, 2022). In order to make further progress in the Narrative Why-Question Answering, one should be knowledgeable about challenges that exist in this domain and in its datasets. These challenges can stem from both the properties of the narrative understanding and the specifics of why-question answering. Some previous works (Lal et al., 2021, 2022) have mentioned commonsense-related challenges. However, we are not aware of any previous work that has attempted to give a comprehensive list of challenges related to this topic.

In this paper, our goal is to give a wider overview of the potential challenges in Narrative Why-Question Answering that can help to inform the researchers working in this domain. Furthermore, in terms of datasets, the TellMeWhy dataset (Lal et al., 2021) is the only dataset that solely focuses on Narrative Why-Question Answering. In this paper, we also address why-questions in multiple-choice, free-form, and extractive narrative QA datasets in order to more fully identify the scope of challenges in this domain. We believe that considering other Narrative Why-Question Answering datasets can also help further development in this domain.

We start by reviewing the concepts and challenges of why-questions and narratives in sections 2 and 3. In section 4, we identify the datasets relevant to Narrative Why-Question Answering, and provide an overview of the commonly used evaluation measures. Finally, in section 5, we analyze the dataset- and task-specific challenges according to the concepts mentioned in previous sections.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 520 - 530

2 Why-Questions

A why-question is typically asked about a causal relation in the text. Causal questions can be constructed in several ways and they are not limited to why-questions only. Causal questions can be also asked with *what* (e.g., what is the cause of), *which* (e.g., which are the consequences of), and *how* (e.g., how dangerous is) (Girju, 2003). However, why-question is the only question type that solely represents causality and can be used to test causal reasoning (Grivaz, 2010; Dunietz et al., 2017; Tan et al., 2022). Furthermore, answering why-questions requires more complex reasoning than answering other types of causal questions because it is more difficult for QA-systems to decide directly from a why-question which type of information needs to be searched for in the text (Girju, 2003). In the following subsections, we start by elaborating on the notion of causality, including how causal relations are expressed in the texts, and then review what ambiguities why-questions hold in relation to causality.

2.1 Causality

Causality is a semantic relationship between events showing that an event occurs or holds due to another event (Mostafazadeh et al., 2016b). Mostafazadeh et al. (2016b) distinguish four types of lexical causality relations: *cause*, *enable*, *prevent*, and *cause-to-end* based on the works by Wolff and Song (2003), Wolff (2007), and Khemlani et al. (2014). Moreover, causality has temporal implications such that if an event A causes/enables/prevents an event B, then A should start before B, or if an event A causes an event B to end, then B should start before A. Causality relations can hold one of the three temporal implications: *before*, *overlaps*, and *during* (Mostafazadeh et al., 2016b). Thus, while answering a why-question, the temporal relation between the events should also be taken into account in addition to the causality relation.

A causal relation is constructed from two components: cause and effect. Based on how the cause and the effect are conveyed in a text, causation can be distinguished into the following categories: explicit vs implicit, marked vs unmarked, and ambiguous vs unambiguous.

Explicit vs Implicit. Causation is explicit if both the cause and the effect are present in the text. Causation is implicit if either the cause or the effect of both are missing from the text (Blanco et al., 2008). For instance, "She was accepted to a top university after receiving a high score in the state examination" is explicit, while "I did not attend the mandatory final exam." is implicit because the effect of "failing the course" is not explicitly stated.

Marked vs Unmarked. Causation is marked if the text contains the causal signal words that indicate the causal relation (Blanco et al., 2008). For example, "I was late *because of* traffic" is marked, but "Do not buy any bread. We have already got two at home" is unmarked.

Ambiguous vs Unambiguous. If the causal relation is presented in the text with causal keywords (e.g., *cause*, *effect*, *consequence*) or with causal signals (e.g., *because of*, *due to*, *as a result of*), it is considered unambiguous (Girju, 2003). On the other hand, if a causal relation is constructed in the form of an expression containing affect verbs (e.g., *affect*, *change*, *influence*) or link verbs (e.g., *link*, *lead*, *depend*), it is considered ambiguous. Furthermore, if a marked signal always refers to causation (e.g., *because*), it is unambiguous, while if a marked word occasionally signals causation (e.g., *since*), it is ambiguous (Blanco et al., 2008).

2.2 Why-Questions: Ambiguity

Why-questions are constructed based on explicit and implicit causal relations in the text. Such questions seek a reason/cause as an answer. However, it is not always clear which reason/cause can be an answer to a question. There are two types of ambiguity: question ambiguity and answer ambiguity.

2.2.1 Question ambiguity

Question ambiguity can occur because of the structural ambiguity in the syntax of the question (Verberne et al., 2006). Due to question ambiguity, it might be not clear what action the why-question refers to. For example, "Why did he say that he will not come to the party?" can be interpreted as "Why did he say it?" or "Why will he not come to the party?". Both "He was asked what he will wear to the party." and "He has other plans for that time." can be correct answers based on different interpretations of the question.

2.2.2 Answer ambiguity

Answer ambiguity occurs because most questions can have multiple answers belonging to different answer types and because often the desired type is not expressed in the question. Several partially

overlapping taxonomies of reasons, which is the cause component of a causal relation, have been proposed (Verberne et al., 2006; Dunietz et al., 2017; Tan et al., 2022). Verberne et al. (2006) distinguish four types of reasons based on Quirk et al. (1985):

- *Cause* - a temporal and causal relation without the involvement of the human intention: an event mechanistically leads to another event;
- *Motivation* - a temporal and causal relation with an involvement of the human intention: a goal or a motivation of an agent leads to their action;
- *Circumstance* - a temporal and causal relation based on conditionality: one event is a condition for another event to occur;
- *Generic purpose* - a causal relation stemming from physical functions of the objects.

Similarly, Dunietz et al. (2017) defines three types of causalities while annotating causal relations: (1) Consequence: similar to the Cause type above, (2) Motivation and (3) Purpose: similar to the Motivation type above. Tan et al. (2022) defines four senses for causality based on Webber et al. (2019) for annotating causal relations: (1) Cause: similar to the Cause type above (2) Purpose: similar to the Motivation type above, (3) Condition and (4) Negative-Condition, which can fit into the Circumstance type above. Although the types of reasons introduced by Verberne et al. (2006) are broader than the taxonomies of Dunietz et al. (2017) and Tan et al. (2022), this list is not complete, as Verberne et al. (2006) demonstrated that not all why-questions can be classified into these categories.

Context: "He opened the box to take a slice of pizza."
Question: "Why did he open the box?"
Answers:
(1) The pizza was in the box.
(2) The box was closed.
(3) He was hungry.
(4) He wanted to eat pizza.
(5) He wanted to take a slice of pizza.

Table 1: An example of answer ambiguity. Answers (1) and (2) refer to causal reasons, answers (3), (4) and (5) refer to motivational reasons.

Valid answers to a why-question about an event or a state can include at least one of the cause, motivation, circumstance, or generic purpose of an event or state according to the above taxonomy. Since a why-question can often be answered with answers falling into several type categories, the necessity to choose the correct answer type creates ambiguity since the desired type is typically not explicitly stated in the question. Furthermore, a why-question can be answered with several causes in the causal chain (Verberne et al., 2006), and in that case all these answers can be considered as correct. For instance, consider the example shown in Table 1. For this example question, several potential causes can be the basis for the answer. Consequently, this why-question can be answered according to both mechanistically causal (answers 1, 2) and motivational (answers 3, 4, 5) reasons.

3 Narratives

Narratives are texts in which events are causally or thematically linked and develop within a temporal framework (Brewer, 2017). Narratives are generally agent-oriented and their main scope is centered on characters, their actions, and motivations (Sang et al., 2022). In narrative QA, stories, fairytales, books, and (movie) scripts are commonly utilized as narrative texts. Characteristics of narrative texts, such as causality of events and motivations of agents, make narratives a suitable context for asking why-questions. Additionally, fictional narratives can ensure the test of comprehension because they are self-contained, meaning that all elements needed to understand the narrative, such as events, characters, and settings, are present in the text and QA models need to comprehend the narrative in order to answer questions (Dunietz et al., 2020; Richardson et al., 2013; Kočiský et al., 2018). Implicitness is a key feature of narratives that makes it different from other types of texts. Length is another characteristic dimension of narratives which is also very important for QA systems. In the following subsections, we will review these characteristics in more detail.

3.1 Implicitness: "Reading between the lines"

People often think and communicate with each other in the form of a narrative (story) (Dunietz et al., 2020). They assume that other people with whom they interact share a common ground with them, so they do not have to mention or specify

commonly known knowledge (Ostermann et al., 2018a). Similar to the implicitness characteristic of the natural narrative-style communication, narrative texts tend to exclude common knowledge, such as commonsense and script (typical sequences of events to accomplish common tasks) knowledge, and assume that the reader has the background knowledge required to infer relevant implicit information (Schank and Abelson, 1975). For instance, not all causes of events and reasons for actions of agents are explicitly stated in narratives. Thus, the ability to "read between the lines" is necessary for properly understanding narratives (Norvig, 1987).

3.2 Short vs long narratives

Narratives can be short or long based on the scope of the text stream and the number of events it contains.

Short narratives cover a small number of events and briefly narrate the actions of fewer agents. The local structure of a longer narrative such as an individual scene can be also considered and used as a short narrative. In short narratives, the reader can make inferences by linking local narrative elements and creating a local narrative representation (Sang et al., 2022; Kintsch, 1988).

Long narratives, on the other hand, have large textual content, cover many events, and focus on the actions and interactions of many agents. Long narratives require the readers to comprehend the underlying deep structure of the narrative and analyze the high-level abstractions. Answering questions in this setting requires understanding the global narrative structure, such as the whole story (Sang et al., 2022; Kintsch, 1988) and the integration of various information stated in different parts of the long narrative by connecting individual scenes (McNamara and Magliano, 2009).

4 Narrative Why-Question Answering

Narrative Why-Question Answering task can be formulated as a special case of Why-QA where the context is a complex structured text—a narrative. Currently, there exists only one QA dataset (TellMeWhy) that solely consists of why-questions in narrative setting. Additionally, several other narrative QA datasets contain why-questions in various proportions. The subsets consisting of why-questions can be extracted from these datasets and used for training or testing Narrative Why-Question Answering systems. In the following subsections, we first review these potential datasets suitable for Narrative Why-QA, and then give an overview of common evaluation measures used to assess the performance of Narrative Why-Question Answering systems.

4.1 Datasets

We selected several multiple-choice, extractive, and free-form QA datasets that utilize narrative as their context. In order to identify why-questions in these datasets, we first extracted all questions including the word *why*. We then manually removed any non-why questions (e.g, "what did the king's son do after he wondered why the girl was crying") from the questions that do not start with *why*. The relevant statistics of all datasets are shown in Table 2.

TellMeWhy (Lal et al., 2021) dataset presents free-form why-questions over events in short narratives. It is the only existing dataset created with the Narrative Why-Question Answering task in mind. The questions were created using template-based transformations and the answers to questions were crowdsourced. Narratives were collected from ROCStories (Mostafazadeh et al., 2016a) and CATERS (Mostafazadeh et al., 2016b). The dataset has a total of 30,519 why-questions with three golden free-form answers for each question. According to data annotators, 28.82% of questions in the dataset cannot be answered explicitly based on the narrative (context).

MCTest (Richardson et al., 2013) is a multiple-choice MRC dataset based on fictional stories. The dataset is created via crowdsourcing and it is designed for the level of understanding of 7-year-old children. The fictional and basic comprehension nature of the dataset decreases the need for additional world knowledge and makes it possible to find the answer only based on the text.

MCScript (Ostermann et al., 2018a) is a multiple-choice MRC dataset based on stories about daily activities. It is created to evaluate machine comprehension using commonsense (script) knowledge (Ostermann et al., 2018b). Stories are collected by crowdsourcing new texts based on selected scenarios. Questions are crowdsourced based on scenarios independent of narratives and then matched with narratives randomly. Similar to MCTest, texts and questions are created according to the understanding level of a child. In general, 27.4% of questions require commonsense (script) knowledge to correctly infer the answer.

Dataset	# of Why	% of Why	Answer	Context	% of Implicit
TellMeWhy	30519	100	free-form	short	28.82
MCTest	329	12.5	multiple-choice	short	-
MCScript	1623	11.6	multiple-choice	short	27.4
MCScript2.0	136	0.6	multiple-choice	short	50
CosmosQA	12439	35	multiple-choice	short	93.8
NarrativeQA	4179	9	free-form	long/summaries	42
FairytaleQA	2864	27	span/free-form	short	25.5

Table 2: Statistics of the narrative why-QA datasets. # of Why shows the number of why-questions in the datasets. % of Why refers to the proportion of why-questions in the datasets. The percentage of implicit questions is taken from the respective dataset papers, except for the NarrativeQA for which this number is due to the analysis done by Bauer et al. (2018)

MCScript2.0 (Ostermann et al., 2019) is another multiple-choice MRC dataset focused on script knowledge. The stories were collected by reusing narratives from the MCScript, and crowdsourcing texts based on new scenarios. Questions were collected based on target sentences of stories rather than scenarios or complete stories. Similar to MCScript and MCTest, the texts and questions are created according to the understanding level of a child. Correct and incorrect answers were crowdsourced by showing questions and hiding the target sentences in the story. In total, 50% of the questions require commonsense knowledge to be answered.

Cosmos QA (Huang et al., 2019) is a multiple-choice commonsense-based reading comprehension dataset. 93.8% of the questions in the dataset require contextual commonsense reasoning. Context paragraphs were collected from the spinn3r blog story corpus Burton et al. (2009) and a dataset by Gordon and Swanson (2009). Both questions and answers were crowdsourced. Questions are based on the causes and effects of events, facts about entities, and counterfactuals.

NarrativeQA (Kočiský et al., 2018) is a narrative reading comprehension dataset based on books and movies. Books from the Project Gutenberg and movie scripts from the web are used as stories. Moreover, summaries for long narratives are obtained from Wikipedia. Questions and answers are crowdsourced based on summaries only. Since both original long stories and summaries exist for each question, this dataset can be used for two tasks: narrative QA based on long narratives (books and movie scripts) and short narratives (summaries). Manual analysis on the validation set by Bauer et al. (2018) showed that 42% of the questions need commonsense knowledge for inference.

FairytaleQA (Xu et al., 2022) is a narrative comprehension dataset designed for both question answering and question generation tasks. The narratives were collected from the Project Gutenberg by considering the reading difficulty up to the 10th-grade level. Small sections were extracted from fairytales as context paragraphs. Following the narrative comprehension frameworks by Paris and Paris (2003) and Alonzo et al. (2009), trained annotators created questions and answers for the contexts. The most common questions are about characters' behavior and causal relationships. 25.5% of the questions are implicit (free-form) and 74.5% of the questions are explicit (span-based).

The amount of why-questions in the reviewed datasets is reported in Table 2. Among the multiple-choice QA datasets, CosmosQA has a higher number of why-questions compared to others. Among the free-form QA datasets, TellMeWhy dataset contains approximately 4.5 times more why-questions than the other two free-form QA datasets combined. Considering the proportion of why-questions in these datasets (also shown in Table 2), why-questions are well-represented in the CosmosQA and FairytaleQA datasets where they make up a sizeable part of the whole dataset, while in the MCTest, MCScript, MCScript2.0, and NarrativeQA datasets, why-questions cover only a small portion of the whole dataset.

4.2 Evaluation measures

For multiple-choice QA datasets, accuracy is a commonly used metric to measure the performance of a model. For free-form QA datasets, both automatic and human evaluation measures are utilized to evaluate the capabilities of the QA model. Most commonly, ROUGE-L (Lin, 2004), Meteor (Denkowski and Lavie, 2011), BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020) and

BertScore (Zhang et al., 2020) have been used to automatically evaluate the performance of the free-form QA models in narrative setting. Overall, F1 score of the ROUGE-L is the most commonly reported automatic evaluation measure.

In terms of human evaluation, Lal et al. (2021) proposed to assess the grammaticality and validity of the answers based on a 5-point Likert scale. The scale of the grammaticality ranges from strongly ungrammatical (1) to strongly grammatical (5), where a strongly grammatical answer must follow all the rules of the English grammar and a neutral score (3) is indicated when the meaning of the answer can be still inferred despite clear grammatical mistakes. The validity scale assesses whether the answer is valid and makes sense in the given context.

5 Challenges

In this section, we list challenges based on our analysis of the datasets and the characteristics of the Narrative Why-Question Answering. This section is divided into three parts. In the first part, we focus on the commonly addressed challenges in the Narrative Why-Question Answering. The existing datasets are created generally to test the abilities of models on these challenges. In the second part, we review some potential challenges that are not marked in existing datasets and that the current models thus cannot be tested on. Lastly, we review challenges that stem from both the datasets and the characteristics of the Narrative Why-Question Answering that can create problems on benchmarking models on this task.

5.1 Commonly focused challenges

In this subsection, we review the challenges that the creators of existing datasets have focused on, related to the implicit vs explicit questions and the length of the narrative.

5.1.1 Explicit vs implicit questions

Questions in the majority of datasets ask about both explicit and implicit causal relations stated in the narratives. The distinction between explicit vs implicit questions is based on the notions stated in sections 2.1 and 3.1:

- *Explicit questions* ask about clearly stated causal relations in the narratives. The answer can be found in the narrative, often as a span of the text. Answering explicit why-questions requires the model to identify affect, link, and causative verbs (e.g., *change*, *lead*, *cause*) or causal signals (e.g., *because of*, *as a result of*, *due to*, *so*) in the narratives (Mirza and Tonelli, 2014).

- *Implicit questions* ask about not explicitly stated causal relations in the narratives. Answering these questions requires filling in the gaps with additional background knowledge, such as commonsense or script knowledge (Norvig, 1987; Lal et al., 2021).

In Table 2, we can see that most reviewed datasets contain more explicit questions than implicit ones. The CosmosQA dataset is an exception here, as it was designed as a commonsense reasoning dataset with a focus on narratives, in which answers cannot be found using the text spans of the context only, and commonsense knowledge is required to answer most questions.

Generally, the inclusion of additional knowledge improves the performance of the QA models to answer implicit questions (Lal et al., 2022). However, in simple stories, such as in MCScript, the system can learn some amount of background knowledge also from the stories and the effect of using additional commonsense knowledge is small (Ostermann et al., 2018b).

5.1.2 Short vs Long narratives

All reviewed datasets have short narratives as their context. The NarrativeQA short texts have a more complex narrative structure than other datasets, since the short context versions of the NarrativeQA are summaries of the larger narratives, and not single scenes from the long narratives. In short narratives, if there is a common lexical pattern between the question and a part of the narrative, or a large lexical overlap between the answer and the narrative, sophisticated models can treat free-form QA as an extractive task. For example, models trained on the TellMeWhy dataset generally try to find the answer span in the text and copy a part of the narrative as an answer (Lal et al., 2021).

The NarrativeQA dataset is the only dataset that has long narratives as its context. Linking narrative elements to answer questions in large narratives is harder than in short narratives (see section 3.2). Typically, in order to reason about long narratives, the parts relevant to reasoning are retrieved first (Kočiský et al., 2018; Tay et al., 2019; Frermann, 2019; Mou et al., 2020, 2021). The retrieval is

difficult even with the state-of-the-art models due to the characteristics of narratives and the necessity of high-level narrative comprehension (Mou et al., 2021).

5.2 Less focused challenges

In this subsection, we list some of the challenges that can be potentially relevant for why-question answering but that the current datasets do not concentrate on. In particular, we review different formats the why-questions can have, and discuss the unmarked and ambiguous causal relations.

5.2.1 Why-question formats

In why-questions, *why* can appear in different parts of the question and questions can be formulated in different ways. Based on the manual analysis of the questions in the datasets listed, we observed the following variations of why-questions.

- *Simple why-question:* the question starts with *why*. Most why-questions in the reviewed datasets follow this format.
- *Long why-question:* the why-question is formulated in a long format such as "What (is/was/may be) (the/a possible/a real) reason why ..." or "(Could/Can) you tell me why ...".
- *Specific why-question:* the question first limits the situation to a certain time/place/person and then formulates the why-question. For instance, "At the end of the story, surrounded by cameras and police, why does Norma think she is on set?" or "According to Bonnie, why is Blanche a constant danger to the gang's well-being?".
- *Statement+why question:* this is a differently formulated type of specific why-question that starts with a statement, which is followed by the one-word question "why?". For example, "When Gandhi was 23, he was thrown off a train in South Africa. Why?"
- *Question chain:* several questions including at least one why-question formulated in one sentence, e.g., "What is Gruul and why are they raiding?".

Most reviewed datasets contain simple why-questions, other variations of why-questions make up a very small portion of the datasets if any. For example, the TellMeWhy dataset contains only simple why-questions since the questions were constructed using question templates where *why* is always the first word of the question. It would be useful to have a fair portion of other why-question formats in the datasets as well in order to test how well the models can handle these other format types. One easy way to accomplish this would be to use templates to transform current simple why-questions into other formats.

5.2.2 Unmarked and ambiguous causality

As discussed in section 2.1, based on causality construction, causation can be categorized to marked vs unmarked and ambiguous vs unambiguous in addition to explicit vs implicit. The reviewed datasets only focus on explicit vs implicit causation nuance (see section 5.1.1) and further categorization is not annotated in these datasets. Thus, it is currently difficult to identify how the models' performance would differ in answering questions asked about marked vs unmarked or ambiguous vs unambiguous causal relations.

5.3 Challenges for benchmarking

In this subsection, we review some challenges that can occur during assessing the performance of models on the datasets. First, we look into what we call the general question problem and list its potential causes. Then, we discuss how the general question problem and the ambiguity of the why-questions can affect the evaluation of models on the Narrative Why-Question Answering task.

5.3.1 General question problem

Tasks designed on narrative QA datasets require systems to answer questions based on narrative (context). Questions, for which a context/narrative is available, should be correctly answerable only based on this context/narrative. If the question can be answered without the narrative, it does not meet the requirements of reading comprehension, especially in narrative QA. We distinguish between the context-specific and general questions as follows:

- *Context-specific questions* are questions that can be answered correctly only by using the information given in the context.
- *General questions* are questions that can be answered without the context.

Although we expect all why-questions in narrative settings to be context-specific, datasets still

contain questions that can be answered both with and without the consideration of the context. This can happen due to the issues in data collection process and due to the characteristics of the causal questions.

If the questions are created without considering the context details, questions can end up being general. This is the case of the MCScript dataset, where questions were asked based on the general scenario description and not on the specific narrative. Thus, some questions that were created are answerable irrespective of the narrative (Ostermann et al., 2019).

When questions have several golden answers (i.e., in free-form QA datasets), a question is considered context-specific if all golden answers can be correctly inferred only by using the information given in the context. In some question-answer pairs, the question can be general because some of the answers do not contain context-specific information. For instance, in the example shown in Table 1, while answers (1), (3), (4), and (5) are specifically related to context, the answer (2) can be correct in any context such as "She opened the box to take out her shoes". A QA model can treat this example question both as context-specific or general, and any of the given five answers can be considered correct. So, in order avoid a question being general, context-specific information can be added to answers that make the question context-specific. In the example of Table 1, "The box was closed." can be converted to "The pizza box was closed." which contains the context-specific word "pizza"; this step makes the question context-specific for all the answers in the given set.

5.3.2 Evaluation

Evaluation of multiple-choice QA datasets. Evaluating answers of why-questions in multiple-choice datasets is a straightforward process. The presence of only one correct answer in multiple-choice questions helps to correctly assess the performance of the model without having to consider the potential answer ambiguity of why-questions. However, if question ambiguity is not addressed in the dataset, some of the questions can have more than one correct answer. Moreover, general why-questions can affect the accuracy of the overall assessment since these questions remove the necessity of the narrative understanding component of the task. Therefore, general questions should be identified and removed from the datasets in order to correctly assess the models' comprehension ability.

Evaluation of free-form QA datasets. General questions can affect the correct evaluation on free-form QA datasets as well. Thus, in order to increase the accuracy of the overall evaluation process, general questions should be identified and transformed to context-specific by adding contextual information to those answers that make the question general. Moreover, ambiguity in why-questions can cause additional problems with both automatic and human evaluations. Due to the question and answer ambiguities, why-questions can have more valid answers than the collected golden answers in the datasets, and collecting all valid answers to these questions is not feasible and is probably impossible. Consequently, automatic metrics can only evaluate the output of models against the set of golden answers, which is likely only a small subset of all valid answers, and thus these metrics cannot fully measure the capacity of the models.

Human evaluation is considered the gold standard in all text generation tasks, including free-form QA (Celikyilmaz et al., 2020). However, performing human evaluation is a costly and slow process (Lal et al., 2021), and the reliability of human judgments is questionable (Gatt and Krahmer, 2018), especially in why-question answering that possesses many ambiguities. For example, human evaluators can prefer one interpretation of the question over another in terms of question ambiguity (section 2.2.1) or consider some causes (e.g., motivational) in the causal chain more reasonable than other causes (e.g., mechanistically causal) in case of answer ambiguity (section 2.2.2). Thus, further instructions are needed in the evaluation process to resolve ambiguities in the why-questions.

6 Conclusion

In this paper, we reviewed challenges and datasets related to Narrative Why-Question Answering. The challenges that occur in this domain can stem from both the properties of narrative understanding and the specifics of why-question answering. The main challenges regarding narrative understanding are the exclusion of common knowledge, the necessity of understanding the local and global narrative structure, and high-level abstraction. The primary challenges of why-question answering are related to identifying causal relations, and ambiguities in questions and answers.

In order to understand data-specific challenges

and the implications of task-specific challenges in datasets, we reviewed seven datasets that can be suitable for this task. We listed challenges that models are tested on in these datasets. The implicitness of questions and the length of narratives are more tested challenges in the datasets. We outlined different why-question formats and questions about unmarked and ambiguous causal relations as other potential challenges that models can be tested on.

Finally, we argued that general questions, and answer and question ambiguities in why-questions can create challenges for benchmarking. We propose that removing general questions and resolving ambiguities in why-questions can lead to more accurate evaluation of systems. We hope that the review of this potential challenges and the analysis of the datasets listed in this paper will help to further the progress in Narrative Why-Question Answering.

Limitations

In the paper, we focus on free-form/span-based and multiple-choice question answering datasets and do not consider other types of QA datasets, such as the cloze test. We also do not focus on datasets that have a mix of narrative and expository contexts. Also, although we took the effort to carefully list the challenges relevant to Narrative Why-Question Answering, it is possible that we missed something; thus, the list of challenges presented in this paper should not be considered complete.

Ethics Statement

In this paper, we review existing datasets and literature related to the task of Narrative Why-Question Answering. The reviewed datasets are publicly available and do not contain sensitive data. We do not publish any new data or experimental results.

Acknowledgements

This work was supported by the Estonian Research Council grant PSG721 and the European Social Fund via the IT Academy programme.

References

Julie Alonzo, Deni Basaraba, Gerald Tindal, and Ronald S. Carriveau. 2009. They read, but how well do they understand?: An empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention*, 35(1):34–44.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Eduardo Blanco, Nuria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1196–1207, New York, NY, USA. Association for Computing Machinery.

William F. Brewer. 2017. Literary theory, rhetoric, and stylistics: Implications for psychology. In *In Theoretical issues in reading comprehension*.

Kevin R. Burton, Akshay Java, and Ian Soboroff. 2009. The icwsm 2009 spinn3r dataset. In *In Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics.

Lea Frermann. 2019. Extractive NarrativeQA with heuristic pre-training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 172–182, Hong Kong, China. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1):65–170.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.

Andrew S. Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *In Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*.

Cécile Grivaz. 2010. Human judgements on causation in French texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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.

Sangeet S. Khemlani, Aron K. Barbey, and Philip N. Johnson-Laird. 2014. Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8.

Walter Kintsch. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics.

Yash Kumar Lal, Horace Liu, Niket Tandon, Nathanael Chambers, Ray Mooney, and Niranjan Balasubramanian. 2022. Analyzing the contribution of commonsense knowledge sources for why-question answering. In *ACL 2022 Workshop on Commonsense Representation and Reasoning*.

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.

Danielle S. McNamara and Joe Magliano. 2009. Chapter 9 toward a comprehensive model of comprehension. In *The Psychology of Learning and Motivation*, volume 51 of *Psychology of Learning and Motivation*, pages 297–384. Academic Press.

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.

Xiangyang Mou, Chenghao Yang, Mo Yu, Bingsheng Yao, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2021. Narrative question answering with cutting-edge open-domain QA techniques: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 9:1032–1046.

Xiangyang Mou, Mo Yu, Bingsheng Yao, Chenghao Yang, Xiaoxiao Guo, Saloni Potdar, and Hui Su. 2020. Frustratingly hard evidence retrieval for QA over books. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 108–113, Online. Association for Computational Linguistics.

Peter Norvig. 1987. *A Unified Theory of Inference for Text Understanding*. Ph.D. thesis, EECS Department, University of California, Berkeley.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018a. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018b. SemEval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 747–757, New Orleans, Louisiana. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension cor-

pus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alison H. Paris and Scott G. Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly*, 38(1):36–76.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive grammar of the English language*. Longman, London ; New York.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Yisi Sang, Xiangyang Mou, Jing Li, Jeffrey Stanton, and Mo Yu. 2022. A survey of machine narrative reading comprehension assessments. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5580–5587. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, page 151–157, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news.

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. 2019. Simple and effective curriculum pointer-generator networks for reading comprehension over long narratives. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4922–4931, Florence, Italy. Association for Computational Linguistics.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2006. Data for question answering: The case of why. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual.

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

Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.