On Narrative Question Answering Skills

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 Question Answering is an important task for evaluating and improving reading comprehension abilities in both humans and machines. However, there is a lack of consensus on the skill taxonomy that would enable systematic and comprehensive assessment and learning of the various aspects of Narrative Question Answering. Existing task-level skill views oversimplify the multidimensional nature of tasks, while question-level taxonomies face issues in evaluation and methodology. To address these challenges, we introduce a more inclusive skill taxonomy that synthesizes and redefines narrative understanding skills from previous taxonomies and includes a generation skill dimension from the answering perspective.

1 Introduction

Narrative Question Answering entails responding to questions based on a narrative context. Understanding narratives requires comprehension of the foundational narrative elements (Zhao et al., 2023) that are not only explicitly stated but also implied in the text, which necessitates "reading between the lines" (Norvig, 1987; Huang et al., 2019). Additionally, answering questions in narrative settings demands forming abstract representations, integrating information across the document in addition to local representation, and generating coherent answers, which may not only be a span of text (Kočiský et al., 2018). Due to requiring unique and multidimensional skills, Narrative Question Answering has become an important task to assess and enhance the various complex reading comprehension abilities of both humans and machines (Dunietz et al., 2020; Sang et al., 2022; Xu et al., 2022). However, a consensus is lacking on the taxonomy of skills that Narrative Question Answering represents and is suitable for assessment (Rogers et al., 2023).

Task-level skill definitions only focus on one characteristic of the task, such as format, and ignore the multidimensional aspect of Narrative Question Answering. In contrast, question-level skill definitions focus on identifying fine-grained skill definitions. Previously, several question-level skill taxonomies have been proposed. Some of these taxonomies (e.g., Sugawara et al., 2017a,b) concentrate solely on challenging reading comprehension skills, such as commonsense reasoning, omitting others. Alternatively, other taxonomies (e.g., Schlegel et al., 2020; Rogers et al., 2023) group skills in a manner that a question can be attributed to several skills within the same dimension, e.g., the question can be associated with both temporal and causal skills from reasoning skill dimension, creating challenges during skill evaluation. Recent narrative reading comprehension taxonomies do not pose these concerns and incorporate clearly distinguishable skills within skill dimensions (Sang et al., 2022; Xu et al., 2022). However, both of the taxonomies focus on a few skill dimensions. Additionally, the taxonomy of Xu et al. (2022) does not clearly define the explicit and implicit skills; thus, adopting this taxonomy might lead to confusion in skill evaluation or in developing methods based on these skill dimensions (e.g., Peng et al., 2023).

We conclude that the current literature lacks a comprehensive Narrative Question Answering skill taxonomy that accurately defines relevant skills without omitting key skills or skill dimensions. To address the identified problems, we introduce a taxonomy that synthesizes skills and skill dimensions from Xu et al. (2022) and Sang et al. (2022), provides accurate definitions for implicit and explicit questions while introducing answer generation skills. Our taxonomy is structured around four skill dimensions: narrative elements, representation scope, knowledge gap filling, and generation. These skill dimensions encompass both understanding and answering skills.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 814–820
June 16-21, 2024

We start by reviewing the previously proposed skill taxonomies and discuss their limitations in Section 2. We introduce our skill taxonomy in Section 3 and consider the skill assessment and learning opportunities of the proposed taxonomy in Section 4.

2 Review of QA Skill Taxonomies

Skills are learned response patterns (Bao et al., 2023) that play a crucial role when answering questions in narrative settings. There are two levels of approach in defining skills for question answering: task-level and question-level skills.

2.1 Task-level skills

Task-level skills are often discussed in multi-tasking setups, where each dataset is treated as a separate skill, or multiple datasets are combined under the same task-level skill (e.g., Khashabi et al., 2020; Zhong et al., 2022; Puerto et al., 2023). Typically, only one characteristic of a task, such as format, is regarded as a skill. However, a key challenge with task-level skills is that the assigned skill to a whole task may not fully represent the entire dataset. For instance, Narrative Question Answering is commonly considered a generative or abstractive task because questions in narrative settings require models to produce answers by generating them based on the information provided in a context (Rogers et al., 2023; Dzendzik et al., 2021; Khashabi et al., 2020). Despite this classification, the task encompasses a wide range of answers, ranging from span-based responses to entirely generative answers that cannot be extracted from the text. Thus, when training on Narrative Question Answering datasets, the model will learn extraction in addition to generation due to the span-based answers in the dataset. Consequently, defining skills at the task level fails to accurately capture the multidimensionality of the task.

2.2 Question-level skills

Question-level skill definitions center on identifying the specific fine-grained skills required to answer each individual question in a dataset. Sugawara et al. (2017a,b) concentrate on general reading comprehension, including Narrative Question Answering, and identified up to 13 prerequisite skills for question answering. However, their focus on challenging skills, like commonsense reasoning, omits considerations for comparably easier skills like recognizing explicit information. This approach to skill taxonomy fails to comprehensively capture the diverse nature of comprehension abilities, posing obstacles to focused and balanced model evaluation. Another notable issue with these taxonomies is that questions may be associated with multiple skills. This introduces an additional challenge in assessing model performance. For instance, if a question requires both temporal and causal reasoning skills, evaluating a model's reasoning ability on these elements might not clearly reveal which skill pattern the model employed to answer the question. This ambiguity could impact the accuracy and specificity of skill assessment in comprehension tasks.

Schlegel et al. (2020) and Rogers et al. (2023) have presented a skill taxonomy based on dimensions. In these frameworks, skill dimensions are orthogonal high-level categories. Each question can be categorized based on at least one of these dimensions and be associated with at least one skill from a dimension, providing a structured way to describe and analyze skills via dimensions. These taxonomies do not omit easier skills, and the orthogonal skill dimension helps to conduct focused evaluation. However, questions may still be attributed to multiple skills from the same dimension, so the ambiguity problem in skill assessment remains unsolved in these taxonomies as well.

Recent studies on narrative reading comprehension also focus on skill dimensions. Sang et al. (2022) introduce meaning representation scope of a narrative and target narrative element skill dimensions for assessment. Xu et al. (2022) uses narrative elements or relations dimension based on Paris and Paris (2003) and source of answers dimension for question annotation schema. Unlike previous taxonomies, these frameworks prioritize narrative elements over reasoning, ensuring that each skill in each dimension is distinguishable. Notably, each question could only correspond to one skill in every dimension, providing clarity in skill attribution. However, both taxonomies concentrate on only two dimensions of Narrative Question Answering skills, leaving other dimensions unaddressed. Furthermore, some of the skill definitions by Xu et al. (2022) are inaccurate. For instance, they define explicit questions as extractive questions and implicit questions as free-form questions requiring high-level summarization. However, the implicit nature of questions should be determined by the information conveyed in the narrative rather than how the answer is constructed or the extent of

the narrative text stream it requires. Inaccurately characterized explicit and implicit questions pose challenges not only in assessment but also in skill-based model development. Due to the definitions proposed by Xu et al. (2022), Peng et al. (2023) treat implicit and generative questions as equivalent concepts, leading them to develop methods based on this inaccurate assumption.

2.3 Summary of Limitations

The task-level skill perspective fails to capture the multidimensionality of the Narrative Question Answering. Previous question-level skill taxonomies either exclude crucial skills or lack distinguishable skills within the skill dimension. While narrative comprehension skill taxonomies (Sang et al., 2022; Xu et al., 2022) address some issues of previous taxonomies, they have been limited to two dimensions. Moreover, one of the taxonomies (Xu et al., 2022) incorrectly defines implicitness skill based on answer format. To better define Narrative Question Answering skills, we synthesize narrative comprehension skill taxonomies, define explicit and implicit questions based on the information conveyed, incorporate high-level summarization as an integral part of the representation scope dimension, and introduce a generation skill dimension.

3 Our Skill Taxonomy

Our skill taxonomy combines elements from prior narrative reading comprehension taxonomies, provides a redefined perspective on implicitness (referred to as knowledge gap filling), and introduces a generation skill dimension. We categorize the skill dimensions into two parts:

- **Understanding Skills**: This involves a model acquiring skills to construct narrative representation and reasoning abilities to answer questions.
- **Answering Skills**: This aspect entails answer formulation skills to effectively represent reasoning over input as an output.

Figure 1 provides an overview of our skill taxonomy, while Figure 2 showcases narrative questions alongside their corresponding skill sets.

3.1 Understanding Skills

3.1.1 Narrative Elements

Narratives center on characters and highlight their actions, interactions, and goals (Graesser et al., 1996; Sang et al., 2022). Questions asked in a narrative context primarily revolve around the narrative elements and relations. Previous taxonomies in general reading comprehension concentrated on the reasoning abilities necessary for extracting narrative elements. Given that extracting a single element or relation might involve multiple reasoning abilities, it becomes challenging during evaluation to discern which reasoning skill patterns the model learned during training. Therefore, a more effective approach is to shift the focus from reasoning to actual individual elements that are inherently more distinguishable. Based on Sang et al. (2022); Xu et al. (2022), our taxonomy contains three narrative elements that narrative questions focus on:

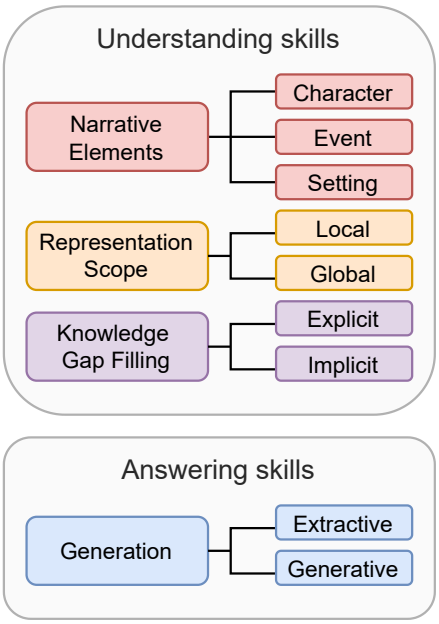

Figure 1: Overview of the proposed skill taxonomy.

- **Character** questions are asked to determine the identity of the characters in the story or describe their characteristics. Questions focus on characters' roles, traits, relationships, emotions, reactions, and facts in the narrative.
- **Event** questions focus on specific events and the actions of characters. Questions ask to identify or predict an event, an action, or a relation between events, such as causal, conditional, and temporal relationships.
- **Setting** questions focus on the specific place, time, and environment in which the events take place. Questions generally ask about where and when the story events happen.

3.1.2 Representation Scope

Forming sufficient narrative representations is vital for successfully comprehending narrative elements and relations (Sang et al., 2022). Representation scope can be defined based on the extent of the narrative text stream required to answer questions (Kintsch, 1988; Sang et al., 2022):

- **Local** narrative representation involves questions related to a single story section, requiring to make local inferences.
- **Global** narrative representation encompasses questions related to multiple story sections, emphasizing the need for high-level summarization.

3.1.3 Knowledge Gap Filling

When communicating, people assume that there is a shared common ground, so information that can be easily recovered is often left unmentioned or underspecified (Ostermann et al., 2018). Since humans use narrative as a core mechanism to think and communicate from childhood (Somasundaran et al., 2018; Dunietz et al., 2020; Piper et al., 2021), narrative texts also inherit these characteristics where common knowledge is omitted. Thus, another skill dimension that the model should succeed in is understanding conveyed information in the narrative and filling in unmentioned information when necessary. This skill dimension can be divided into:

- **Explicit** questions are those for which the information required to answer them is directly and clearly presented in the text. These questions typically pertain to facts, details, or events that are explicitly mentioned in the narrative. In other words, the answer to an explicit question can be found by referring to the information given in the text itself.
- **Implicit** questions are those that require readers to infer or deduce information that is not explicitly stated in the text. Answering implicit questions often necessitates the use of commonsense knowledge and the ability to "read between the lines" (Norvig, 1987). Implicit questions often involve understanding causation, identifying relationships, or making inferences about events or reasons that are not directly provided in the narrative (Huang et al., 2019; Lal et al., 2021; Kalbaliyev and Sirts, 2022).

3.2 Answering Skills

3.2.1 Generation

A crucial skill in Narrative Question Answering involves effectively representing reasoning through answer generation, particularly concerning how the reader formulates an answer based on the question and the narrative.

In everyday communication, individuals often repeat names or lengthy phrases in conversation (Gu et al., 2016). Similarly, when responding to questions, answers can vary from straightforward span-based responses to entirely generative answers that cannot be directly copied from the narrative. Hence, it is essential for the reader to learn the distinction between reusing the semantic concepts given in the narrative and selecting new semantic concepts from the reader's vocabulary to construct an answer. Since copying and generating operate at the concept level, we differentiate question-level skills by categorizing questions as either extractive or generative based on the structure of the answers in answer formulation:

- **Extractive** questions require answers that exist as spans within the narrative and can be extracted and formulated from the narrative.
- **Generative** questions necessitate answers that cannot be solely constructed by extracting and formulating spans from the narrative. Instead, they require additional words or phrases to either complement the extracted span or form the complete answer.

4 Skill Assessment and Learning

Effective and fair skill evaluations rely heavily on precisely defining and annotating the dimensions and features under study. In past instances, conducting focused assessments and fair evaluations posed challenges due to combining multiple skill dimensions into one and overlapping numerous skill features. Our taxonomy outlined in Section 3 addresses these issues by distinguishing each skill dimension and its associated features. Annotating existing datasets and constructing future datasets based on our taxonomy will assist in conducting focused assessments and ensuring fair skill evaluations. The distinguishability of skill dimensions allows the study of individual dimensions in isolation for focused assessment. For fair skill evaluation, we assert that every skill feature within each dimension holds equal significance, and a single skill

Narrative:

[...] the snowflakes, as they fell upon Thumbelina, were like a whole shovelful falling upon one of us, for we are tall, but she was only an inch high.
[…]
She came at last to the door of a field mouse, who had a little den under the corn stubble […] Poor Thumbelina stood before the door, just like a little beggar girl, and asked for a small piece of corn […]
[...]
"You poor little creature," said the field mouse, for she was really a good old mouse, "come into my warm room and dine with me." She was pleased with Thumbelina, so she said, "You are quite welcome to stay with me all the winter, if you like" […] Thumbelina […] found herself very comfortable […]

Questions, answers, and skills:

Q_1: How tall is Thumbelina?
A_1: an inch high
Skills: character, local, explicit, extractive

Q_2: Why did Thumbelina feel comfortable in the den?
A_2: the field mouse gave her corn to eat and a place to stay.
Skills: event, global, implicit, generative

Q_3: Where did Thumbelina live in January?
A_3: in the den of the field mouse
Skills: setting, global, implicit, extractive

Figure 2: Examples of narrative questions and associated skills. Color and underlining emphasize the main concepts in the narrative that are related to those in the answer and the question, respectively. *Note: The example narrative is an excerpt from "Little Thumbelina" by Hans Christian Andersen and taken from Project Gutenberg. This narrative has also been used in the FairytaleQA (Xu et al., 2022) dataset.*

feature should not be the only representation of the whole task or dataset. Instead of presenting results as a singular dataset-level or task-level metric, we advocate for showcasing results across various skill dimensions and features. This approach provides a more comprehensive understanding of the model's performance and contributes to a more accurate and fair evaluation of skills. However, as Narrative Question Answering is free-form in nature, challenges related to the evaluation of text generation (Celikyilmaz et al., 2021) also persist in skill evaluation, making it essential to consider that automatic measures might not fully demonstrate models' abilities.

In terms of skill learning, each skill dimension becomes a focal point for improvement. Previously, methods have been developed for enhancing input representations with narrative elements (e.g., Bao et al., 2023; Peng et al., 2023), utilizing knowledge-based methods to enhance implicitness (e.g., Bauer et al., 2018; Lal et al., 2022), and employing Pointer Generator Networks (See et al., 2017) to improve the generation ability of models (e.g., Bauer et al., 2018; Tay et al., 2019; Nishida et al., 2019; Peng et al., 2023). It is crucial to note that some of these previous efforts (e.g., Peng et al., 2023) have relied on inaccurate definitions of skills. We anticipate enhanced performance in Narrative Question Answering by considering model development and annotation based on accurate skills definitions. Furthermore, existing skill-based methods often concentrate on improving a single dimension or even a specific skill. We argue that there is considerable room for improvement by shifting the focus to multiple skill dimensions during method development. We believe assigning equal significance to each skill feature throughout the development process is key to achieving more robust and effective results.

5 Conclusion

Narrative Question Answering is a crucial task to assess and enhance complex reading comprehension skills. However, there is no consensus regarding the classification of skills that Narrative Question Answering entails and whether they are suitable for evaluation. The current research lacks a comprehensive taxonomy that contains and correctly defines relevant essential skills and skill dimensions. In this paper, we propose a skill taxonomy for Narrative Question Answering to address these challenges. Our taxonomy synthesizes and redefines skills from existing taxonomies while incorporating a generation skill dimension. Our taxonomy contains distinguishable skills within narrative elements, representation scope, knowledge gap filling, and generation skill dimensions. We hope that our taxonomy will facilitate focused and fair multidimensional skills assessment of Narrative Question Answering and motivate further development of skill-learning methods.

Limitations

We focus on Narrative Question Answering skills that make questions distinguishable within a dimension for fair evaluation. Thus, we do not consider

other skill dimensions, such as linguistic skills, that do not fit our criteria. We specifically concentrate on Narrative Question Answering, which is free-form in nature. Although some of the skill dimensions are applicable to other narrative comprehension tasks, we do not consider specific skills of other tasks. Our references are primarily from the studies conducted in English; however, the skills identified in both previous and our studies are applicable across all languages.

Acknowledgements

This work was supported by the Estonian Research Council Grant PSG721.

References

Meikai Bao, Qi Liu, Kai Zhang, Ye Liu, Linan Yue, Longfei Li, and Jun Zhou. 2023. Keep skills in mind: Understanding and implementing skills in commonsense question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5012–5020. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey.

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.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English machine reading comprehension datasets: A survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arthur Graesser, Jonathan M. Golding, and Debra L. Long. 1996. Narrative Representation and Comprehension. In Rebecca Barr, Michael L. Kamil, Peter B. Mosenthal, and P. David Pearson, editors, *Handbook of Reading Research, Volume II*, pages 171–205. Routledge.

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

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.

Emil Kalbaliyev and Kairit Sirts. 2022. Narrative why-question answering: A review of challenges and datasets. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 520–530, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why-questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kyosuke Nishida, Itsumi Saito, Kosuke Nishida, Kazutoshi Shinoda, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2019. Multi-style generative reading comprehension. In *Proceedings of the 57th Annual*