To Test Machine Comprehension, Start by Defining Comprehension

Jesse Dunietz,[*] **Gregory Burnham,**[*] **Akash Bharadwaj, Owen Rambow,**
Jennifer Chu-Carroll, and **David Ferrucci**
Elemental Cognition
{jessed,gregb,akashb,owenr,jenniferc,davef}
@elementalcognition.com

Abstract

Many tasks aim to measure MACHINE READING COMPREHENSION (MRC), often focusing on question types presumed to be difficult. Rarely, however, do task designers start by considering what systems should in fact comprehend. In this paper we make two key contributions. First, we argue that existing approaches do not adequately define comprehension; they are too unsystematic about what content is tested. Second, we present a detailed definition of comprehension—a TEMPLATE OF UNDERSTANDING—for a widely useful class of texts, namely short narratives. We then conduct an experiment that strongly suggests existing systems are not up to the task of narrative understanding as we define it.

1 Introduction

Over the past few years, neural models (e.g., Chen et al., 2016; Devlin et al., 2019; Liu et al., 2019) have begun to match or even exceed human performance on MACHINE READING COMPREHENSION (MRC) benchmarks. In these tasks, systems demonstrate their comprehension of a passage by answering questions about it. Yet despite recent successes, MRC appears far from solved: systems continue to make basic, sometimes baffling mistakes, and they fail to generalize to new data. Such shortcomings have motivated a flurry of new MRC tasks, each designed to confront systems with questions deemed challenging for current methods. For example, tasks may ask questions requiring commonsense reasoning (Huang et al., 2019), multi-hop reasoning (Welbl et al., 2018), or inferences based on a second passage (Lin et al., 2019).

This line of research assumes that ever-more-"difficult" question-answering tasks will ultimately lead to more robust and useful reading comprehension. We argue that, while the question-answering

format can be a fine choice for **how** to test comprehension, using difficulty as the basis for **what** to test is fundamentally flawed. To put it provocatively, the dominant MRC research paradigm is like trying to become a professional sprinter by glancing around the gym and adopting any exercises that look hard. The training may end up exercising some relevant muscles, but it is far too haphazard to achieve the ultimate goal.

Like athletic training, MRC tasks are not an end in themselves; ultimately, they are meant to lead to real-world applications. Current tasks may suffice for sufficiently similar applications—e.g., chatbots that look up customer questions in product documentation. But many proposed NLP applications hinge on deeper comprehension. Early work (e.g., Dyer, 1982) pointed to examples like assistance with legal disputes and service contracts; more recent work suggests applications such as summarizing a patient's clinical timeline (Jung et al., 2011). For such complex applications, machines will need to manipulate rich models of the world evoked by the text—e.g., to compare a claimant's narrative to legal standards, or to build a causal model of a patient's condition. From this broader perspective, the current paradigm falls short.

Specifically, we claim that in the quest for difficulty, task designers overlook the issue of what **content**—what information expressed, implied, or relied on by the passage—systems should comprehend. MRC datasets are usually constructed by having humans cast about for supposedly tricky questions, most often questions based on reasoning. But the questions that result are scattershot, offering little assurance that even a high-scoring system has achieved a useful and robust understanding.

We advocate for a different approach. We propose that the first step in defining MRC tasks should be specifying what content a system would likely need to understand for a given class of applica-

---

[*]Equal contributions.

tions. Only then can tasks systematically compile questions to probe for the internal model that the machine ought to have constructed.

This paper demonstrates such an approach for applications that involve understanding narratives.[1] After reviewing existing approaches to constructing MRC datasets (§2), we argue for narratives as a valuable MRC testbed (§3.1). Then, inspired by cognitive science research on reading comprehension, we propose a "template of understanding" (ToU) for stories—an account of what an internal model of a story should minimally contain (§3.2). We also suggest ways to operationalize our ToU as a story comprehension task (§4). Finally, we show evidence from a pilot ToU-based task that current MRC models are not up to the challenge (§5).

2 Existing MRC dataset designs

This paper addresses how MRC tests can be made more systematic. Accordingly, we review existing tasks grouped by their data collection methods. We argue that each category falls short of testing a useful body of content in a satisfying way.

2.1 Manually written questions

By far the most popular strategy for generating MRC questions is to have humans—usually crowd workers, but sometimes trained annotators—think of questions about each passage.

The most straightforward version of this method gives annotators little to no guidance regarding what questions to ask. One early example is the TREC-8 dataset (Voorhees and Tice, 2000). In the more recent SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) entailment tasks, the only constraint on crowd workers was that they produce one entailed, one contradicted, and one neutral hypothesis for each premise sentence.[2] Similarly, the workers who assembled NewsQA (Trischler et al., 2017) were told only that the questions had to be answerable with short phrases, and workers for SQuAD (Rajpurkar et al., 2016) were simply given a "good" and a "bad" example and encouraged to use original wording.

The problem with such an open-ended generation process is that, absent stronger guidance, people tend to write simple questions that can be answered using lexical cues. (See, e.g., the dataset analysis in Rajpurkar et al., 2016.) This makes the tasks questionable measures of comprehension.

The dominant solution is to incorporate trickier twists. NarrativeQA (Kočiský et al., 2018) and DuoRC (Saha et al., 2018) reduce lexical similarity between questions and passages by showing annotators only a second passage about the same events. Other datasets emphasize reasoning presumed to be difficult, such as incorporating information from multiple parts of the text. MCTest (Richardson et al., 2013) and MultiRC (Khashabi et al., 2018) ask for questions that rely on multiple sentences; ROPES (Lin et al., 2019) has annotators apply information from one passage to write questions on a second; and HotpotQA (Yang et al., 2018b) and QASC (Khot et al., 2019) require multi-hop reasoning. Other forms of reasoning tested include coreference resolution (Quoref, Dasigi et al., 2019; Winograd Schema Challange, Levesque et al., 2012), numerical reasoning (DROP, Dua et al., 2019), and commonsense reasoning (Cosmos QA, Huang et al., 2019). Tasks can also be made harder with devices such as unanswerable questions (SQuADRUn, Rajpurkar et al., 2018; NewsQA; CosmosQA) and filtering questions with an adversarial baseline (DROP; Quoref; QASC).

These twists do make MRC harder. But to pursue hard questions is to overlook why easy questions seemed inadequate in the first place: MRC tasks are a means to an end, namely useful applications, and easy questions—e.g., questions that depend only on lexical cues—do not suffice for that end. The techniques above may help by guiding annotators to a different space of questions: intuition suggests that some of these harder questions are indeed useful ones. But such techniques are an incomplete solution, as difficulty is a weak proxy for utility. What matters is not the system's sophistication per se; it is the alignment between the questions the system can answer and the ones a given application needs it to. Designing for difficulty still gives little assurance of such alignment.

Perhaps a truly random walk through question space would eventually cover a representative set of useful questions, but annotators are biased toward questions that humans find interesting (see Gordon and Van Durme, 2013; Misra et al., 2016; Zhang

---

[1] We will use "narrative" and "story" interchangeably, roughly following the Wikipedia definition: "A narrative or story is an account of a series of related events, experiences, or the like, whether true... or fictitious."

[2] Parts of the original RTE datasets (Dagan et al., 2006, etc.) were generated more systematically, but only in the sense that the outputs of NLP tools (e.g., translation or information extraction systems) were recorded as correct/incorrect examples of entailment. Little attention was paid to subject matter.

et al., 2017). They do not think to ask questions whose answers seem obvious, even when those answers are essential to comprehension. If we do not delineate such facts and evaluate systems' ability to manipulate them, we will never be satisfied that the systems have adequately understood the text.

2.2 Naturally occurring questions

A second approach is to find questions "in the wild," then retrospectively collect documents containing the answers. This is the approach of BoolQ (Clark et al., 2019) and MS MARCO (Nguyen et al., 2016), which compile search engine queries, and of ELI5 (Fan et al., 2019), which harvests questions from Reddit's "Explain Like I'm Five" forum.

Such datasets are clearly useful for answering common queries, a valuable application class in its own right. For more complex applications, however, common queries are, if anything, less thorough than annotators at probing important elements of understanding (particularly aspects humans find obvious). The mismatch between questions and passage content is exacerbated by finding the passages retrospectively: the questions do not even attempt to test most of what each passage discusses, making them an insufficient measure of MRC.

2.3 Questions from tests designed for humans

The third strategy is to pull questions from tests written for humans. Examples include the early "Deep Read" corpus (Hirschman et al., 1999); the more recent TriviaQA (Joshi et al., 2017) and SearchQA (Dunn et al., 2017) datasets, which mine collections of trivia questions; the AI2 Reasoning Challenge (ARC; Clark et al., 2018), which asks questions from standardized science tests; and RACE (Lai et al., 2017), which draws from English learning materials for Chinese school students.

Our chief concern about this approach echoes our concerns from §2.1: tests designed for humans rarely bother to test content that most humans find obvious. Accordingly, they gloss over vast swaths of understanding that machines do not yet have but which may be critical to applications. In addition, SearchQA, TriviaQA, and ARC find passages retrospectively, so again, the questions they ask only tangentially graze the content of each passage.

2.4 Automatically generated questions

Several projects generate questions algorithmically. The *CNN/Daily Mail* datasets (Hermann et al., 2015) and ReCoRD (Zhang et al., 2018) produce cloze-style questions over news passages by masking out entities from summaries and below-the-fold sentences. ComplexWebQuestions (CWQ; Talmor and Berant, 2018) and WikiHop (Welbl et al., 2018) test for multi-hop reasoning by walking a structured knowledge base. Finally, bAbI (Weston et al., 2016) generates short texts and questions from a simple simulation of characters moving around.

Each algorithm encodes assumptions about what is worth asking. In theory, then, the algorithmic approach could produce a satisfying MRC test: given appropriate inputs, the algorithm could aim to generate questions that cover important content. Indeed, our proposal in §4.1 can be seen as a question generation algorithm to be run by humans.

In practice, however, algorithmic approaches have de-emphasized content. *CNN/Daily Mail* and ReCoRD capture explicit assertions about maskable entities, which do not amount to a principled body of content. The algorithms behind CWQ and WikiHop at least take as input some body of content, namely knowledge graphs. But the graphs include only a fraction—again, not a principled one—of the associated documents' content, and the questions are further restricted to rely on multi-hop reasoning. Multi-hop reasoning is no doubt a major error source for MRC, but applications are driven by what propositions must be extracted; whether each proposition takes zero inference steps or seven is immaterial. Accordingly, multi-hop questions are worth investigating, but they are not a sufficiently well-motivated body of content to constitute a measure of reading comprehension.

Similar remarks can be made about most of bAbI's 20 "tasks": grounded in simulations, their question generation algorithms start from known content, but target forms of reasoning. However, the tasks concerning time, positions, sizes, pathfinding, and motivations are closer to our content-first question generation strategy. These tasks are not driven by applications, and their synthetic passages are unrealistically simple, but among existing datasets, they are closest to our proposal.

2.5 Summary: What is missing

The most clear-cut way to test reading comprehension would be to select passages, describe what should be comprehended from them, and design tests for that understanding. Yet few MRC datasets have even approximated this approach. Many impose little structure on what content is tested; the

rest pick some "difficult" form(s) of analysis or linguistic phenomena, but rarely consider downstream goals to determine what the questions should be **about**. Metrics for difficult reasoning and linguistic phenomena (see, e.g., Gardner et al., 2019) are useful, but only as tools for error analysis and mitigation; they are not top-line performance metrics.

In addition, many datasets to date suffer from two other problems: 1) they select passages after the questions are asked, meaning the questions test comprehension of only small portions of the passages; and/or 2) they ask very few questions whose answers are obvious to humans.

These issues of content scope also intersect with issues of format. Many tasks have adopted a span extraction format, including TREC QA, NewsQA, and (most notably) SQuAD and its successors. This format immediately rules out questions about inferred events or entities, which may be essential to a complete interpretation. The main alternative is multiple choice (MC), used in tasks such as Cosmos QA, RACE, ARC, WikiHop, and every task in GLUE (Wang et al., 2019b) and SuperGLUE (Wang et al., 2019a). But MC has its own problem of providing extra hints via answer choices.

We will return to the format issue in §4. But first, we propose a more systematic approach to constructing MRC datasets.

3 Defining deep story understanding

Our approach starts from the **content** of a passage, which we define as the information it expresses, implies, or relies on. Specifically, we propose that task designers lay out a minimal body of content that MRC systems should demonstrate they understand. Exactly what that content is will vary from passage to passage, of course, but the key is to define a TEMPLATE OF UNDERSTANDING (ToU): a set of question templates that can be filled in with specific events and entities for any given passage. The answers to the fleshed-out questions will constitute a floor of understanding for the passage—a plausible lower bound on what content machines ought to comprehend.

The natural next question is what content the ToU should cover. System needs will vary by application. To advance MRC writ large without limiting ourselves to a single application, we propose selecting a class of texts where one could reasonably predict a priori what content would be useful for applications. In the rest of this section, we endorse

fictional narratives as a particularly promising class of texts and propose a ToU for them.[3]

3.1 The case for stories

Stories have several convenient properties that recommend them as a testbed for MRC.

Most importantly, applications that involve comprehending stories are numerous and diverse. Consider a legal aid tool: to assess whether a lawsuit may be warranted, it would have to comprehend an account of the events in question. Likewise, a tool that finds candidates for medical trials would need to read each patient history. (Appendix A fleshes out these scenarios.) These examples are not exceptional; applications in other domains will depend on stories in customer complaints, intelligence dispatches, financial news, and many other document types. Humans tend to think and communicate in terms of stories (see, e.g., Haidt, 2013; Mateas and Sengers, 1999; Bruner, 1991; Eck, 2006), so it is unsurprising that stories are ubiquitous in the content we want NLU tools to help us with.

Additionally, stories come with a strong prior from cognitive science about what elements of understanding will be useful. Research on human reading comprehension (e.g., Graesser et al., 1994; Zwaan et al., 1995) suggests that humans attend primarily to the timeline of events, to the locations of entities and events, and to the causes and motivations of events and actions. For applications that involve story comprehension, we can expect that machines will need to understand these same dimensions. We can thus design a principled ToU for stories even without specifying an application.

Stories' content also makes them a particularly compelling demonstration of understanding, for two reasons. First, cognitive science suggests that humans make more inferences when reading narrative text than expository text (Graesser et al., 1994). In particular, a story entails a highly structured network of relations (timelines, causality, etc.). Thus, stories do exercise abilities beyond simple factoid extraction. Second, stories rely on a large body of implicit world knowledge. If a system is able to use and express that knowledge when reading stories, it will likely be able to apply the same knowledge even when comprehending other kinds of texts.

Among stories, fictional ones offer the strongest test of comprehension: their contents cannot be

---

[3]To be clear, we are **not** claiming that fictional narratives are themselves an application; only that they are a class of texts that are useful for many applications.

found in corpora, so systems must rely on comprehending the text (Richardson et al., 2013). Accordingly, we suggest using fictional narratives as the basis for developing a ToU and evaluating MRC.

3.2 A ToU for stories

We propose four overlapping clusters of questions for story comprehension, corresponding to the four elements identified by Zwaan et al. (1995) as the ones humans attend to when reading stories. Further support for these questions, particularly the last two, comes from early work in computational story understanding: Schank and Abelson (1977) identify causal chains, plans and goals as crucial elements of understanding multi-sentence stories.

1. **Spatial:** Where are entities positioned over time, relative to landmarks and each other? How are they physically oriented? And where do events take place?

2. **Temporal:** What events and sub-events occur, and in what order? Also, for what blocks of that timeline do entities' states hold true?

3. **Causal:** How do events and states lead mechanistically to the events and states described or implied by the text?

4. **Motivational:** How do agents' beliefs, desires, and emotions lead to their actions?

These question templates form the ToU. Systems should ideally be able to answer them about all entities and events that the story mentions or implies (though of course some entities/events are more important than others; see §4.1). We do not have a separate category for "who did what to whom" information, but we expect strong performance on the ToU to hinge on such analysis. In particular, much of this information is captured in the characterization of events for temporal questions.

Of course, these four facets do not cover everything one might comprehend. They include nothing about the story's message, or how it resembles other stories, or even most counting questions. The ToU merely provides a lower bound on what is needed. That said, many forms of reasoning (e.g., counting) can be reduced to deterministically manipulating the answers to multiple ToU questions.

4 Towards a story understanding task

Our ToU provides a conceptual framework for stating what a machine should understand from a story.

Spatial (*sample entries*):
- Rover is in the yard from when he runs out the door until he runs inside.
- Rover is in the house from when he runs inside until the end of the story.

Temporal (*sample entries*):
- Allie arrives just before Rover runs outside.
- Rover barks just before he runs inside.
- It is still raining at the end of the story.

Motivational (*sample entry*):
- Rover runs inside, rather than staying put, because:
  - If he runs inside, he will be inside, whereas if he does not he will be outside, because:
    * Rover is outside.
    * Running to a place results in being there.
  - If Rover is inside, he will not get rained on, whereas if he is outside he will, because:
    * It is raining.
    * When it is raining, things that are outside tend to get rained on, whereas things inside do not.
  - Rover would prefer not getting rained on to getting rained on, because:
    * Most dogs prefer not to get rained on.

Figure 1: A partial RoU for the following simple story fragment: . . . *One day, it was raining. When Allie arrived, Rover ran out the door. He barked when he felt the rain. He ran right back inside.*

However, there remains the challenge of operationalizing the framework—i.e., of rigorously assessing whether a machine has that understanding.

We do not claim to have solved this problem, but in this section we discuss two broad directions for further development: evaluating based on annotated answers to ToU questions and asking untrained humans to rank different answers. These approaches might even be combined to offer complementary perspectives on system performance.

4.1 Approach 1: Annotating ToU answers

One class of approaches starts with trained annotators writing plain-English answers to each ToU question. The annotators are given guidelines for instantiating the ToU on new stories and for making answers detailed and thorough. We call an annotator's answer document a RECORD OF UNDERSTANDING (RoU); see Figure 1 for an example.

Conceptually, answering temporal and spatial questions is straightforward, but the causal and motivational questions require more definition. People accept many kinds of answers to such questions. It is therefore important to clarify what a good answer should include—i.e., what causal or motivational

facts an MRC system should comprehend.

We base our account of these questions on the philosophical literature on causality (see Schaffer, 2016) and on the social science literature on what explanations people seek (see Miller, 2019). Following this scholarship, we conceptualize a causal or motivational question as asking what root cause led the event or state from the story to happen rather than some alternative outcome. For example, in a story about Rover the dog, the question of why Rover came inside is taken to mean: *Why did Rover come inside, rather than remaining where he was?*[4]

The answer to such a question is a CAUSAL CHAIN tracing from the root cause to the event or state described in the story (see Figure 2 for examples). The links in the chain walk in lock-step through two parallel worlds: the REALIZED WORLD, where the root cause held true and led to the observed outcome; and an ALTERNATIVE WORLD, where the root cause would have been changed and led to some alternative outcome.

For mechanistic causation, each link in the chain ends in an event that helped bring about the outcome described in the story. For example, two mechanistic links from Figure 2a are *the plant looks brown (rather than green) because it is unhealthy (rather than healthy)* and *the plant is unhealthy because it has little light (rather than lots of light)*.

For motivations, the structure is slightly different. Rather than the final link being an event that happened in the story, it is a statement of the agent's preferences (in Figure 2b, *Rover would prefer not being rained on to being rained on*). The links leading to it are the future causes and effects that the agent imagines will lead from their action to their preferred outcome (e.g., going inside leading to being inside leading to not getting rained on).

The causal chain provides the backbone of an explanation for an event or action, but the full explanation should recursively explain each link (e.g., *Rover would prefer not being rained on to being rained on*). Recursive explanations appeal to some combination of general knowledge about the world (e.g., *Most dogs prefer not to get rained on*) and

story-specific SUPPORTING FACTS—e.g., the fact that Rover is outside. Supporting facts generally need to be recursively explained, as well.

Even with guidelines, different annotators may give substantively different answers. In particular, they may drill down to different levels of detail in a causal chain before bottoming out in general knowledge—e.g., rather than stopping at dogs disliking rain, one annotator might explain that Rover disprefers rain because he dislikes getting wet, which in turn is because dogs often dislike getting wet. To handle such disagreements, we can adopt the pyramid method (Nenkova and Passonneau, 2004) from abstractive summarization, another task where annotators may provide different but equally sensible ground truths. Under this method, a reconciler merges RoUs into a single rubric by identifying shared content "nuggets" (e.g., that it is raining) and weighting each by how many annotators cited it. (See Voorhees [2004] for more on nuggets.)

4.1.1 Preliminary notes on RoU agreement

We conducted a small pilot study on RoU annotation: with the help of 5 annotators, we iteratively crafted guidelines and tested them on 12 stories. Here we share some initial qualitative observations.

For spatial annotations, agreement improved when annotators first drew a simple sketch of each scene, then translated their sketches into statements. This process seemed to help annotators notice implicit spatial facts. Some annotators also reported that sketches lowered the cognitive burden.

For temporal annotations, annotators generally agreed on what events took place and the temporal relations between them. Disagreements stemmed mainly from choices of which implicit occurrences to annotate. We are exploring ways to promote consistency, including having annotators draw timelines to draw attention to missing events. We are also looking to incorporate prior art (e.g., TimeML; Pustejovsky et al., 2003) into our guidelines.

On causal and motivational questions, we were pleasantly surprised by the conceptual consistency between annotators. Annotators appealed to similar causal assertions, even bottoming out in similarly detailed general rules. What was less consistent was structure—how causal chains were carved into links and how bullets were nested. Annotators also occasionally omitted self-evident general rules or supporting facts. We are optimistic that both issues can be improved by more examples and training.

As expected, annotators occasionally differed on

---

[4]Causality as contrast may seem unintuitive, particularly since "why" questions tend not to state a contrasting outcome. But the audience generally just infers a reasonable default.

Beyond its support in the literature, contrast offers several advantages. It makes it far easier to match intuitions about what should factor into a causal explanation. It also naturally handles relative preferences, and allows explaining multiple aspects of an event—e.g., John walking carefully can be explained in contrast to both staying put and walking normally.

	the plant is in the bedroom	the plant has insufficient light	the plant is unhealthy	the plant is brown
Realized world	the plant is in the bedroom →	the plant has insufficient light →	the plant is unhealthy →	the plant is brown
	vs.	*vs.*	*vs.*	*vs.*
Alternative world	the plant is somewhere well-lit	the plant has sufficient light	the plant is healthy	the plant is green

(a) A mechanistic causal chain for the question, "Why did the plant turn brown?"

Realized world	Rover runs in	Rover is inside	Rover does not get rained on	Rover is more satisfied
	vs. →	*vs.* →	*vs.* →	*vs.*
Alternative world	Rover stays put	Rover is outside	Rover gets rained on	Rover is less satisfied

(b) A motivational causal chain for the question, "Why did Rover the dog run back inside when it started raining?"
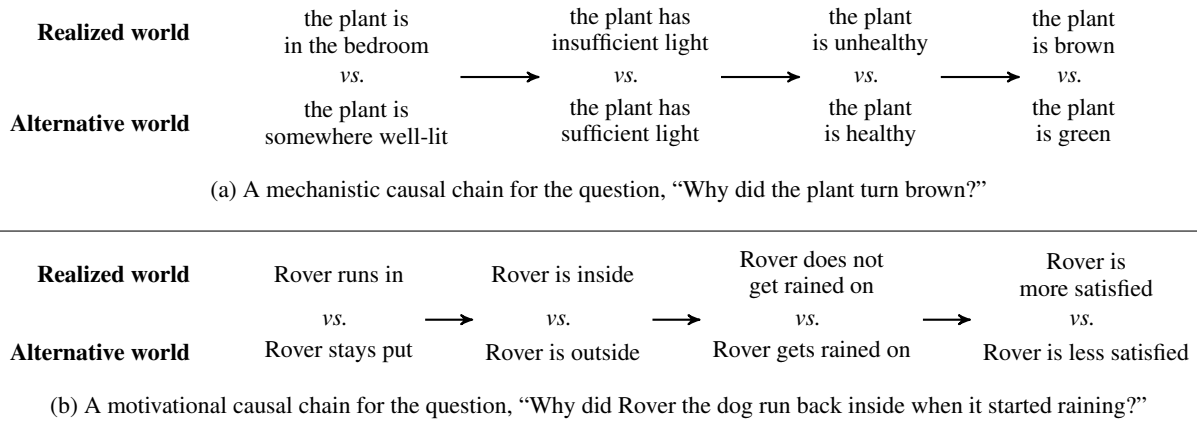
Figure 2: Example causal chains answering causal (above) and motivational (below) ToU questions.

which causal contrasts to include. Such borderline judgments of salience may be inevitable, and seem to warrant use of the pyramid method.

4.1.2 Free-text evaluation

It is difficult to evaluate a system directly on an RoU or a rubric, as they are written in plain English. One option is to pose broad ToU questions (e.g., "What events happened and in what order?") and then to automatically compare systems' full free-text answers to annotators'. But this would require an automated comparison metric, and existing metrics such as ROUGE and BLEU are concerned only with lexical similarity. Their correlation with humans' quality judgments is substantial but not stellar (Callison-Burch et al., 2006), and high scores do not always indicate good answers in MRC (see Yang et al., 2018a; Nema and Khapra, 2018). Superficial similarity measures may prove particularly weak given how open-ended ToU questions are.

Alternatively, human evaluators could read both the RoU-derived rubric and the system output and decide whether the output adequately covers each nugget from the rubric. This is how the pyramid method is typically applied in summarization.

Still a third possibility is to have human evaluators ask targeted questions about each nugget from the rubric. The evaluators could then judge whether the system's shorter free-text answers reflect a consistent understanding of that nugget. Such evaluation would be especially powerful if the evaluators knew the NLP systems' typical shortcuts and could reword a given question accordingly: a suspicious evaluator could query for the same fact in multiple ways to verify that the system consistently gets it right. This would make results more satisfying than many MRC evaluations, as systems couldn't rely on terse answers being interpreted charitably.

Of course, using humans for the final evaluation is expensive, even if automated metrics are used during model development. Human evaluators also add variability and subjectivity, as they may probe differently for the same knowledge or find a given answer more or less convincing. Still, new tasks often start with human evaluation while the community fine-tunes what is worth measuring, and only later to progress to automated metrics that approximate human judgment. Such were the trajectories of topic model coherence (see Lau et al., 2014), summarization (see Yang et al., 2016), and machine translation (see Papineni et al., 2002), so it is a plausible pathway for RoU evaluation, too.

4.1.3 Thorough multiple-choice evaluation

Free-response is a compelling format that is tricky to evaluate. Multiple-choice inverts the trade-off: it is less compelling, but much easier to evaluate.

With the help of the ToU, a multiple-choice (MC) test can be fairly comprehensive. Question writers would first write out RoUs for a story, and perhaps reconcile them into a weighted rubric. They would then write MC questions targeting each nugget in the rubric: What goal is Rover pursuing by running inside rather than staying put? Where was Rover after he ran through the door? How were Rover, the house, and the rain positioned at the end of the story? Etc. Such a thorough MC test based on RoUs would be a step up from current tasks.

The downside of an MC task is that, though easy to evaluate, it would be questionable as a measure of comprehension. All MC tasks suffer from the same lack of naturalness: questions do not normally come with candidate answers, and ranking candidates is simply easier than the tasks MRC

should ultimately support. Furthermore, systems learn to exploit incidental surface features in the question, sometimes performing well even without seeing the passage (Kaushik and Lipton, 2018). When humans take MC tests, we can make strong assumptions about what they must know or do to succeed; an NLP system offers no such assurances.

In the long run, then, we do not see multiple choice as an adequate format for demonstrating MRC. Still, such tests offer some leverage for progress in the short term.

4.2 Approach 2: Competing to satisfy judges

The RoU guidelines put a stake in the ground as to how ToU questions should be answered. But as noted above, ToU questions, particularly "why" questions, admit many good answers. The ones canonicalized by the guidelines and by annotators following them may not always be the most useful.

Consequently, it may prove beneficial to appeal directly to human intuition about what understanding entails. We have assumed that what lets humans perform story-related tasks is that they possess some internal answers to the ToU. If we further assume that humans can be led to favor machine answers that resemble their own internal ones, then humans should make good judges of answer quality even without the guidance of RoUs.

Accordingly, we could let humans judge system's full free-text answers based only on intuitive preferences. Evaluators could still be guided to ask ToU questions thoroughly, but extensive guidelines would not be needed: neither asking questions nor recognizing good answers demands nearly as much specification as stating canonical answers.

Whereas the approaches in §4.1 must strive for replicability in humans' answers, this approach seeks replicability only in humans' **judgments** of answers. We suggest two ways to achieve this.

First, in the absence of a rubric, we suspect that answers would best be judged via pairwise comparisons. For free-text writing, humans generally find comparative assessment easier than absolute scoring (Pollitt, 2012), and comparison is already used to evaluate natural-language generation (see, e.g., Yatskar et al., 2014). Comparisons also mitigate the difficulty of spotting errors of omission: when evaluators see an incomplete answer in isolation, they may gloss over or mentally fill in what was left unsaid. Comparing against a more complete competing answer makes it easier to notice gaps.

Second, evaluators can be guided to tease apart their judgments into several desirable dimensions of explanations—e.g., accuracy, depth, and coherence—just as is often done for natural language generation. Pilot studies would be required to refine the dimensions and their specifications.

5 Current MRC systems do not comprehend stories

If current systems performed well on the ToU, our argument would be moot. This section presents evidence that they do not.

5.1 Data and experimental setup

To test existing systems, the questions must be presented in a form the systems can handle. Many systems were designed for span extraction, but the ToU does not lend itself to answering with text spans. Instead, we report on experiments with a pilot version of the MC task described in §4.1.3.

To construct the test, we selected the first two narrative stories in the dev set of RACE (Lai et al., 2017). Based on our preliminary annotation guidelines, one annotator read both stories, drafted an RoU for each, and wrote a question for each statement in the rough RoUs. The annotator then collaborated with several others to write distractor answers, each characterized by one or more of the following: small surface variations on the correct answer that change the meaning; language from the passage, especially words that appear near words from the question; and language that might plausibly collocate with words from the question.

As an additional test for robustness, questions came in "variant groups": each question was paired with a variant, or occasionally more than one, that asks for the same information in a different way (see Figure 3). The distractors were often altered as well. We then evaluated accuracy in two ways: counting each question independently and counting each variant group as one unit. In the latter method, the group is marked correct only if both variants were answered correctly. This simulates a suspicious evaluator re-asking the question and deducting points if the model does not consistently exhibit the desired understanding.

The resulting dataset contains a total of 201 questions (98 variant groups). 29% are spatial or temporal; the remaining 71% are causal or motivational. The questions average 5.1 options, with a minimum of 4. (Including many distractors somewhat

Figure 3: An example variant group from our ToU-based questions; correct answers *in italics*. In the associated RACE story, a man tricks Mr. Green into driving him home under the pretense of guiding Mr. Green to a hotel. See Appendix B for the full story text.

mitigates the weaknesses of the MC format.) All questions are included in the supplementary materials; Appendix B shows many examples.

For validation, the questions were presented to two colleagues with non-technical degrees. They scored 96% and 91% (measured on variant groups), suggesting that motivated, well-educated humans have little trouble with our questions.

Finally, we put the questions to XLNet (Yang et al., 2019),[5] a large, transformer-based language model trained with generalized autoregression on BooksCorpus and English Wikipedia. After fine-tuning, the model achieves 81.75% on the original RACE task (within 5 points of the best non-ensemble model at the time of the experiments).

5.2 Results and Discussion

Our results (Table 1) show that XLNet performs poorly. On individual questions, it scores just 37%, closing less than a third of the gap between chance and human performance. This strongly suggests that whatever XLNet is doing, it is not learning the ToU's crucial elements of world understanding. Furthermore, the system's performance is brittle, with many correct answers attributable to luck and/or unreliable cues: when moving from questions to variant groups, human performance falls just 3 points. XLNet's performance, on the other

---

[5]For questions with more than four answers, we split the answers across multiple sub-questions, all of whose answer sets contained the correct answer. We counted the question correct only if that answer was chosen across all answer sets. Chance performance was adjusted accordingly.

	All	By question type	
		Spatial + Temporal	Causal + Motivational
Per question	37%	33%	38%
Chance	15%	20%	13%
Human (avg.)	96%	93%	97%
Per variant group	20%	14%	23%
Chance	4%	5%	5%
Human (avg.)	93%	90%	95%

Table 1: XLNet accuracy on our ToU-based questions.

hand, falls 17 points, which leaves the system closing just 18% of the chance-vs.-human gap.

Although we tested only XLNet, all the other models that currently dominate the leaderboards are similar pre-trained language models; none has any distinguishing characteristic that might be expected to produce dramatically better results on our dataset. Likewise, no existing dataset is so much more systematic than RACE that fine-tuning on it should dramatically improve results on our dataset. Especially given that multiple-choice tests are artificially easy for systems (see §4.1.3), our pilot experiment offers strong evidence that existing MRC systems do not succeed on the ToU.

6 Taking the ToU idea forward

Our ToU for stories is a first attempt at defining what MRC systems should comprehend in a principled, systematic way. Drawing on work in psychology, philosophy, and pedagogy, we have argued for the ToU as a minimal standard and a valuable target for MRC. We have also shown it to be beyond the reach of current systems.

We therefore suggest that the NLP community further build on our ToU. This includes refining and perhaps expanding the questions; better defining the answers and evaluation procedures; building MRC corpora based on the ToU; and developing better-performing systems. We ourselves are working on all four, and we welcome collaboration.

But even beyond our ToU, the broader point stands: existing MRC approaches are not satisfactorily testing for a systematic set of content. Our efforts demonstrate that it is possible, with a sufficiently interdisciplinary approach, to define a plausible floor for comprehension for a given class of applications. If MRC is to achieve its ultimate goals, we—the NLP community—owe it to ourselves to ensure that our reading comprehension tests actually test for the comprehension we desire.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Jerome Bruner. 1991. The narrative construction of reality. *Critical Inquiry*, 18(1):1–21.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/ Daily Mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *CoRR*, arXiv:1803.05457v1 [cs.AI].

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential 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 5927–5934, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *CoRR*, arXiv:1704.05179v3 [cs.CL].

Michael George Dyer. 1982. *In-depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*. Ph.D. thesis, Yale University, New Haven, CT, USA. AAI8220836.

Jill Eck. 2006. An analysis of the effectiveness of storytelling with adult learners in supervisory management. Master's thesis, University of Wisconsin–Stout, Menomonie, WI.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor, and Sewon Min. 2019. On making reading comprehension more comprehensive. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 105–112, Hong Kong, China. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3):371.

Jonathan Haidt. 2013. *The Righteous Mind: Why Good People are Divided by Politics and Religion*, chapter 12. Vintage Books.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep Read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332, College Park, Maryland, USA. 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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Hyuckchul Jung, James Allen, Nate Blaylock, Will de Beaumont, Lucian Galescu, and Mary Swift. 2011. Building timelines from narrative clinical records: Initial results based on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 146–154, Stroudsburg, PA, USA. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2019. QASC: A dataset for question answering via sentence composition. *CoRR*, arXiv:1910.11473 [cs.CL].

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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561. AAAI Press.

Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning over paragraph effects in situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 58–62, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, arXiv:1907.11692v1 [cs.CL].

Michael Mateas and Phoebe Sengers. 1999. Narrative intelligence. In *Narrative Intelligence: Papers from the 1999 Fall Symposium*, pages 1–10, Stanford, CA, USA.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, arXiv:1611.09268v3 [cs.CL].

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.

Alastair Pollitt. 2012. Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2):157–170.

James Pustejovsky, José M Castaño, Robert Ingria, Roser Saurí, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New directions in question answering: Papers from the AAAI Spring Symposium*, volume 3, pages 28–34, Stanford, CA, USA.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Schaffer. 2016. The metaphysics of causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2016 edition. Metaphysics Research Lab, Stanford University.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Ellen M. Voorhees. 2004. Overview of the TREC 2003 question answering track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)*, pages 54–68. National Institute of Standards and Technology.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pages 200–207, New York, NY, USA. ACM.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, LA, USA. OpenReview.net.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In

4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li. 2018a. Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 98–104, Melbourne, Australia. Association for Computational Linguistics.

Qian Yang, Rebecca J Passonneau, and Gerard De Melo. 2016. PEAK: Pyramid evaluation via automated knowledge extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2673–2680. AAAI Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018b. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *CoRR*, arXiv:1810.12885v1 [cs.CL].

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Rolf A. Zwaan, Mark C. Langston, and Arthur C. Graesser. 1995. The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5):292–297.

A Examples of applying the ToU to stories for applications

In the main text (§3.1), we suggested that many advanced applications hinge on understanding the elements captured by our ToU for stories. Here we offer several examples from two domains.

A.1 Law

For the foreseeable future, legal decision-making will be the province of lawyers, not AI. However, one plausible use for MRC in a legal setting is as a **screening tool** for helping non-lawyers determine whether a case has enough merit to bother bringing in a lawyer.

For example, consider the first-person narrative below (fictional, but based on an amalgam of several real news stories):

> My property borders on public lands where hunting is allowed. Last month, a hunter tracked a buck onto my property. He claims he didn't see my boundary sign. He ended up stepping up onto the remains of an old stone wall, which crumbled, and he broke his wrist. Now he's saying I can give him $10K now and he'll walk away, or else he's going to sue me for much more.

Before contracting a lawyer, the property owner may want to assess whether there is any merit to the threat. On the other side of the deal, a law firm that offers free initial consultations may wish to avoid wasting time on cases that are clear non-starters.

A second legal application for NLU tools might be helping a lawyer **search for precedents**. For instance, a tool could help with the narrative above (or perhaps a third-person version of it) by looking for cases with similar elements—e.g., an accidental trespass resulting in injury.

To assist in such application scenarios, a system would of course need information about legal codes. But it would also have to understand what happened in the cases it is trying to analyze. To that end, the answers to ToU questions would be essential, as demonstrated in Table 2. The table shows ToU questions and answers that would be key to understanding the landowners situation. (These questions are ones the system would answer for itself while reading, not necessarily questions it would be asked by a user.)

A.2 Medicine

Medicine also offers ample opportunity for an MRC system competent in the narrative ToU to assist doctors and researchers. Narratives pervade electronic health records in the form of doctors' notes, which record information ranging from patient history to detailed descriptions of surgical procedures.

One narrative-based medical application is helping doctors **understand a prior doctor's rationale**. Currently, doctors often spend time sifting through a patient's records to understand why a prior doctor made a certain decision. The reasoning is often explained, but many documents must be searched to find the relevant note.

For example, consider the real medical note below,[6] recorded after a routine follow-up appointment following breast cancer treatment:

> She underwent radiation treatment ending in May 2008. She then started on Arimidex, but unfortunately she did not tolerate the Arimidex and I changed her to Femara. She also did not tolerate the Femara and I changed it to tamoxifen. She did not tolerate the tamoxifen and therefore when I saw her on 11/23/09, she decided that she would take no further antiestrogen therapy. She met with me again on 02/22/10, and decided she wants to rechallenge herself with tamoxifen. When I saw her on 04/28/10, she was really doing quite well with tamoxifen. She tells me 2 weeks after that visit, she developed toxicity from the tamoxifen and therefore stopped it herself. She is not going take to any further tamoxifen.

A future doctor may wonder why the patient is not on hormone therapy, which would be standard procedure. This explanatory note may be hard to find amongst the many notes in the patient's record.

A second medical application is **finding patients who qualify for medical trials**. For instance, a pharmaceutical company might develop a new anti-estrogen drug that they believe has milder side effects. They would then want to find patients who had already tried several anti-estrogen drugs,

---

[6]Quoted from `https://www.mtsamples.com/site/pages/sample.asp?Type=96-&Sample=1939-Breast%20Cancer%20Followup%20-%201`

Question type	ToU question	Example (partial) answer to ToU question	Significance to legal application
Spatial	Where was the hunter when he broke his wrist?	On the landowner's property.	The locations of events are legally relevant in many ways. For one, property owners may be held liable for injuries that occur on their property. Additionally, however, property owners may not be liable for injuries suffered by trespassers.
Spatial	Where was the boundary sign?	On the boundary between the public lands and the writer's property.	The presence of a sign may shield the landowner from responsibility, but recognizing that means understanding that it would mark the boundary between the two properties.
Temporal	When did the stone wall fall into disrepair?	Sometime before the story started.	How long the wall has been in disrepair may be legally relevant. Since the exact timing was not given, the system might flag this question for further clarification.
Temporal	Has the hunter sued?	No, although he may do so in the future.	If the hunter had already sued, the landowner might need representation whether or not the suit had merit.
Causal	Why did the hunter break his wrist (rather than his wrist remaining intact)?	Because he stepped onto the wall (rather than stepping elsewhere), which led to him falling (rather than remaining upright, because the wall was in disrepair rather than better condition), which led to him breaking his wrist (rather than his wrist remaining intact).	The wall's disrepair was allegedly an important causal factor in the injury, making it more plausible that the landowner could be held responsible.
Motivational	Why did the hunter claim he didn't see a sign (rather than saying nothing of signs)?	He would prefer that others believe that he entered the property unwittingly (rather than deliberately), either because he in fact enter unwittingly or because he would like to deny his deliberate violation. He believes that if he says he did not see a sign, others will be more likely to believe this (whereas if he says nothing, they may assume he saw the sign).	The hunter's claim of unwitting entry could be motivated either by true innocence or by deception, which affects whether it should be believed—and unwitting entry may be treated differently by the law. The system may want to flag this claim for follow-up questions about its plausibility.
Causal	Why did the hunter enter the private property (rather than stopping at the boundary)?	Possibly because the hunter didn't see the sign (rather than seeing it), so he remained unaware he was crossing the boundary (rather than realizing he was).	There may be a mechanistic (non-motivational) explanation for why the hunter did not stop at the boundary, and again, unintentional entry may be legally different. Also, the landowner may have been responsible for posting signs that would keep people away from his property if there were any hazards.
Motivational (recursive explanation for the end of the previous causal chain)	Why might being aware of the boundary have made the hunter stop, whereas being unaware of it (may have) led him to cross it?	The hunter likely prefers staying within the law to violating it. If he had known he was at the boundary of private property, he would have known that continuing past the boundary would be illegal trespass, but not knowing about the boundary meant he did not know continuing could be trespassing.	The hunter suggested that missing the sign led to accidentally entering the property, but that claim hinges on the assumption that had he known about the property line, he would have respected it. That may be a challengeable assumption.
Motivational	Why did the hunter threaten to sue, rather than suing immediately?	The hunter would prefer to get less money than to possibly get more money but experience the hassle of a lawsuit and risk getting nothing. He believed that if he threatened, the property owner might be afraid of losing more money and give him the $10,000 (whereas if the hunter sued immediately he would have no chance to avoid the hassle and risk).	It is possible that the very act of extorting money via a threat of a lawsuit has legal implications. Also, this action by the hunter may indicate that he considers the risk of losing the case high or that he is otherwise reluctant to pursue a lawsuit, which may affect what course of action the landowner ultimately wants to take.

Table 2: Example ToU questions and answers for a legal application.

Question type	ToU question	Example (partial) answer to ToU question	Significance to medical application
Temporal	When did the patient start and stop taking tamoxifen?	Multiple times: She started taking it sometime after May 2008 and stopped taking it by 11/23/09. Then, she started taking it again on 02/22/10, and stopped taking it by mid-May 2010.	A clinical trial may be seeking patients who kept stopping and starting a specific drug. It may also be important how long the side effects took to develop. Also note that if the question of interest is really a counting question (how many times), this relies most of all on an underlying temporal understanding like the one captured by the ToU.
Causal/ Motivational	Why is the patient not taking an anti-estrogen drug (rather than taking one)?	She was taking Arimidex, and it caused strong side effects (rather than her having mild or no side effects). Preferring fewer side effects, she therefore tried Femara (rather than continuing with Arimidex). Femara also caused side effects, so for the same reasons as before, she tried switching to tamoxifen (rather than continuing the Femara), but it also caused side effects. The patient preferred not experiencing the side effects to having the medical benefits of the drugs, so she decided not to take any such drug (rather than continuing with one of the above).	A future doctor may expect the patient to be on an anti-estrogen drug, as that is standard for someone with her history of breast cancer. Understanding that the patient has tried many drugs and decided to stop them may inform the doctors course of action. The doctor might proceed differently if he determined that she had stopped for some other reasone.g., that she simply lapsed in a prescription. Also, a clinical trial may be seeking patients who stopped taking a drug because of side effects. Furthermore, the trial might be seeking specifically patients who stopped taking the drug at the advice of the doctor.

Table 3: Example ToU questions and answers for a medical application.

perhaps multiple times, and had toxicity problems with all of them. Currently, research hospitals find patients for a given clinical trial by employing humans to read through the hospital's database of medical notes and determine which patients meet the trial's criteria.

To assist in such application scenarios, an automated system would have to understand medical notes like the one above. In the rationale-finding application, it would have to interpret the note well enough to recognize that it explains the current medical regimen; in the patient-finding application, the system would have to recognize that this patient went on and off of several anti-estrogen drugs because of side effects. Again, understanding the answers to ToU questions would be essential, as demonstrated in Table 3.

B Example ToU-based multiple-choice questions on a RACE story

B.1 The story

Mr. Green was traveling around the country in his car. One evening he was driving along a road and looking for a small hotel when he saw an old man at the side of the road. He stopped his car and said to the old man, "I want to go to the Sun Hotel. Do you know it?"

"Yes." The old man answered. "I'll show you

the way."

He got into Mr. Green's car and they drove for about twelve miles. When they came to a small house, the old man said, "Stop here."

Mr. Green stopped and looked at the house. "But this isn't a hotel." He said to the old man.

"No," the old man answered, "This is my house. And now I'll show you the way to the Sun Hotel. Turn around and go back nine miles. Then you'll see the Sun Hotel on the left."

B.2 RACE's original questions

Answers marked correct by RACE are italicized.

Q1. Where did Mr. Green want to sleep that night?
A) In his car.
B) In his own house.
C) *In a hotel.*
D) In the old man's house.

Q2. Why did Mr. Green stop his car?
A) Because he found a hotel.
B) Because the lights were red.
C) *Because he saw an old man.*
D) Because he saw a friend.

Q3. Where did the old man promise to take Mr. Green?
  A) To Mr. Green's house.
  B) To the old man's house.
  C) To the SunHotel. [sic]
  D) To the country.

Q4. Why didn't the old man stop Mr. Green when they passed the hotel?
  A) Because he wanted Mr. Green to sleep in his house.
  B) Because he wanted to get home.
  C) Because he didn't see the hotel.
  D) Because he didn't know the hotel.

Q5. How far was it from the place where Mr. Green met the old man to the Sun Hotel?
  A) About nine miles.
  B) About three miles.
  C) About twenty-one miles.
  D) About twelve miles.

B.3 A sampling of our ToU-based questions

Correct answers are italicized. Questions are numbered with the IDs used in our dataset, which is available in this paper's supplementary data. The first number in each question ID indicates the variant group; the second number is a group-independent question index.

B.3.1 Causal chains

The questions below target different parts of causal chains explaining why the agents in the story took the actions that they did. The first five ask about why Mr. Green stopped his car (vs. continuing to drive); the next five ask about why the old man said he would show Mr. Green the way (vs. just giving him directions).

Q1-1. Why did Mr. Green stop his car the first time?
  A) Because if he stopped his car, he could ask the man something.
  B) Because if he stopped his car, he could make a new friend.
  C) Because if he stopped his car, the old man could get in.
  D) Because the directions he asked for said to stop the car.
  E) Because if he stopped his car, he could drive for about twelve more miles.
  F) Because he got a flat tire.
  G) Because he was driving along a road.
  H) Because he was traveling around the country.
  I) Because he said, "I want to go to the Sun Hotel".

Q2-3. Why did Mr. Green want to ask the man something?
  A) Because there was something he didn't know.
  B) Because he liked to ask questions.
  C) Because there was a chance to make a friend.
  D) Because he didn't want to drive past the man without helping him.
  E) Because if he stopped his car, he could drive for about twelve miles.
  F) Because he got a flat tire.
  G) Because he was driving along a road.
  H) Because he said, "I want to go to the Sun Hotel".

Q3-7. Before they spoke at all, what did Mr. Green hope the man would be able to do?
  A) Tell him where the hotel was.
  B) Tell him where the small house was.
  C) Get in his car.
  D) Drive for about twelve miles.
  E) Take him to his house.
  F) Take him to the hotel.
  G) See an old man.

Q4-9. What did Mr. Green hope the conversation with the old man would enable him to do?
  A) Get where he was going
  B) Travel around the country
  C) See what he was seeing
  D) Stop and look at a house
  E) Drive with the old man
  F) Come to a small house
  G) Turn around and go back

Q5-11. What was Mr. Green trying to do throughout the story?

  A) *To stay at the small hotel*
  B) To drive along a road
  C) To pass the small hotel
  D) To come to a small house
  E) To see the old man
  F) To stop at the side of the road
  G) To speak with the old man

Q6-12. Why did the old man make his initial offer to Mr. Green?

  A) *The old man was appearing to help Mr. Green while actually tricking him.*
  B) The old man was appearing to trick Mr. Green while actually helping him.
  C) Mr. Green was appearing to help the old man while actually tricking him.
  D) Mr. Green was appearing to trick the old man while actually helping him.

Q7-14. Why did the old man say he would show Mr. Green the way instead of just giving directions?

  A) *So Mr. Green would let him into his car.*
  B) So Mr. Green would stop his car.
  C) So Mr. Green would say something to the old man.
  D) So he could answer Mr. Green.
  E) So they could go to the hotel.
  F) So Mr. Green would take him to the hotel.

Q10-20. Where did the old man expect he and Mr. Green would drive together to?

  A) *The house*
  B) The Sun Hotel
  C) The side of the road
  D) Back nine miles

Q11-22. Why did the man want to ride with Mr. Green?

  A) *He wanted to get home.*
  B) He wanted to get to the hotel.
  C) He wanted to stand at the side of the road.
  D) He wanted to answer Mr. Green.
  E) He wanted to get into Mr. Green's car.

Q13-26. What is one reason the man's plan worked?

  A) *Mr. Green wouldn't know where they were really going.*
  B) Mr. Green wouldn't know what his name really was.
  C) Mr. Green wouldn't know how old he really was.
  D) He wanted to see the hotel on the left.
  E) He showed Mr. Green the way to the hotel.

B.3.2 General knowledge

For causal and motivational questions, an RoU often includes abstract general knowledge. To interrogate these components of understanding, we we wrote questions where the answer choices do not mention any of the entities in the story. Below are general knowledge questions that target the same two events as the questions immediately above.

While we thought these questions might be especially difficult, XLNet handled them about as well as the causal/motivational questions whose answer choices explicitly mentioned story entities.

Q21-44. What is part of the reason why Mr. Green stopped driving when he first saw the man?

  A) *In order to ask someone a question, you have to be close to them.*
  B) In order to get where you're going, you need to stop your car.
  C) When you travel around the country, you stop your car.
  D) When the evening arrives, you drive your car home.
  E) When you're looking for a hotel, you often stop your car.
  F) People often pick up hitchhikers.
  G) People often stop to help others.

Q22-47. Why did Mr. Green think the man on the side of the road might be able to help him?

  A) *Often a person in a given area is familiar with the geography of that area.*
  B) Often a person in a given area gives out useful items.
  C) Often one person can give a ride to another person.
  D) Often a person on the side of the road needs help.

Q23-48. Why did Mr. Green want to know where the hotel was?

  A) *Getting to a place usually requires knowing where the place is.*
  B) Driving around the country usually requires knowing where you are.
  C) Talking with a person usually requires seeing where they are.
  D) Getting into a car usually requires knowing where the car is.

Q24-51. Why was Mr. Green seeking the old man's help in the first place?

  A) *People like to sleep comfortably at night.*
  B) People like to travel in a leisurely manner around the country.
  C) People like to talk amiably with each other.
  D) People like to see interesting sights on the road.
  E) People like to be driven directly to their homes.

Q25-52. Why did the old man say he would show Mr. Green the way, the first time?

  A) *People sometimes trick others for their own gain.*
  B) People sometimes trick others in order to help them.
  C) People sometimes help others for selfless reasons.
  D) People sometimes help others for selfish reasons.

Q26-54. Why did the old man first say he would show Mr. Green the way instead of just giving directions?

  A) *To show someone the way means going along with them whereas giving directions means just telling them information.*
  B) To show someone the way means just giving them information whereas giving directions means going along with them.
  C) Giving directions is more effective than showing someone the way.
  D) Giving directions is less effective than showing someone the way.
  E) Giving directions is more friendly than showing someone the way.
  F) Giving directions is less friendly than showing someone the way.

Q28-58. Why did the old man expect to be able to control the route as he rode with Mr. Green?

  A) *When taking directions, people generally go where they are told to go.*
  B) When taking directions, people usually go somewhere other than where they are told to go.
  C) When on vacation, people generally follow their itineraries.
  D) When driving with strangers, people are generally very careful.
  E) When going to a small house, people generally ride together.

Q29-60. What helps explain why the man wanted to accompany Mr. Green on his drive?

  A) *People usually want to go home at night.*
  B) People usually want to go to a hotel at night.
  C) People usually want to travel around the country.
  D) People usually want to drive with each other.

Q30-62. Why did the old man trick Mr. Green?

  A) *Being driven home by someone is nice and convenient.*
  B) Traveling around the country with someone is fun and exciting.
  C) Stopping and looking at someone's house is interesting and enjoyable.
  D) Answering someone's questions is fulfilling and helpful.

Q31-64. What is one reason the man's plan worked?

  A) *If someone is unfamiliar with an area, they won't realize if they're going the wrong way.*
  B) If someone is familiar with an area, they won't realize if they're going the wrong way.
  C) If someone is unfamiliar with an area, they will realize if they're going the wrong way.
  D) If someone is traveling around the country by car, they will drive an old man's home.
  E) If someone wants to go to a hotel, they will go to a small house first.

B.3.3 Spatio-temporal questions

The questions below target the spatial and temporal information in the story, asking how things were physically arranged at different points in time.

Q37-76. Who was in the car at first?

A) *Mr. Green*
B) Both Mr. Green and the old man
C) The old man
D) Neither Mr. Green nor the old man

Q38-78. Who was in the car when Mr. Green drove to the small house?

A) *Both Mr. Green and the old man*
B) Mr. Green
C) The old man
D) Neither Mr. Green nor the old man

Q39-80. Who was probably in the car when Mr. Green drove away from the small house?

A) *Mr. Green*
B) Both Mr. Green and the old man
C) The old man
D) Neither Mr. Green nor the old man

Q40-82. Who was at the small house at first?

A) *Neither Mr. Green nor the old man*
B) Mr. Green
C) Both Mr. Green and the old man
D) The old man

Q41-84. Who was at the small house when Mr. Green arrived there?

A) *Both Mr. Green and the old man*
B) Mr. Green
C) The old man
D) Neither Mr. Green nor the old man

Q42-86. Who was likely at the small house a short while after the story ends?

A) *The old man*
B) Mr. Green
C) Both Mr. Green and the old man
D) Neither Mr. Green nor the old man

Q53-109. When driving to the old man's, on which side did they pass the hotel?

A) *The car passed the hotel on the right side of the road*
B) The car passed the hotel on the left side of the road
C) The car passed the house on the left side of the road
D) The car passed the house on the right side of the road

Q54-111. How were Mr. Green, the car, the old man, and the window probably situated when Mr. Green stopped to ask the man a question?

A) *Mr. Green in the car, the window down, the man on the side of the road*
B) Mr. Green in the car, the window down, the man in the car
C) Mr. Green in the car, the window up, the man on the side of the road
D) Mr. Green in the car, the window up, the man in the car
E) Mr. Green out of the car, the window down, the man in the car
F) Mr. Green out of the car, the window up, the man in the car

Q55-113. While the two men drove to the old man's house, how was the scene likely arranged?

A) *Mr. Green and the man next to each other, in the car*
B) The man next to Mr. Green next to the car
C) The car in the man and Mr. Green
D) Mr. Green next to the man next to the car
E) The man at his house and Mr. Green in the car
F) Mr. Green at the hotel and the man at his house
G) Mr. Green at his house and the man at the hotel

Q56-115. When Mr. Green was actually going the right way at the end, how was the scene likely arranged?

A) *The man at his house and Mr. Green in the car*
B) Mr. Green and the man next to each other, in the car
C) The man next to Mr. Green next to the car
D) The car in the man and Mr. Green
E) Mr. Green next to the man next to the car
F) Mr. Green at the hotel and the man at his house
G) Mr. Green at his house and the man at the hotel

B.3.4   More variant groups

As described in the paper, for each question we wrote a second version that targeted essentially the same information in a different way. Below are additional examples of such variant groups.

Q19-39. Why could the man still help Mr. Green by showing him the way at the end of the story?

A) *Mr. Green still didn't know how to get to the hotel.*

B) Mr. Green still didn't know that he was at the man's house.

C) Mr. Green was still looking at the house.

D) The old man knew where Mr. Green's car was.

Q19-40. What information was Mr. Green missing that the man provided when he showed him the way the second time?

A) *Mr. Green didn't know how to get to the hotel.*

B) Mr. Green didn't know that he was at the old man's house.

C) Mr. Green didn't know who the old man was.

D) The old man knew where Mr. Green's car was.

Q46-94. Who was in the car just before Mr. Green met the old man?

A) *Mr. Green*

B) Both Mr. Green and the old man

C) The old man

D) Neither Mr. Green nor the old man

Q46-95. Who was in the car when Mr. Green approached the spot where he met the old man?

A) *Mr. Green*

B) Both Mr. Green and the old man

C) The old man

D) Neither Mr. Green nor the old man

Q22-45. Why did Mr. Green want to speak to the old man?

A) *People ask questions when they lack information.*

B) People are interested in the places they travel.

C) People are often very curious.

D) Old men at the side of the road sometimes know the future.

E) People ask questions before letting people into their cars.

F) People interrogate hitchhikers before picking them up.

Q22-46. Why did Mr. Green think the old man might be able to help him?

A) *Sometimes one person has information another person doesn't.*

B) Sometimes one person trades a car for another person's house.

C) Sometimes one person gives a ride to another person.

D) Sometimes one person on the side of the road gets in another person's car.