Accessible Abstract: Why do we answer questions? Sometimes it’s to provide information, which has been the interpretation of the computer science community. But sometimes it’s to probe or test intelligence. This paper argues we should think more about that application of question answering and its connection to the foundations of artificial intelligence: The Turing Test. We thus argue that in addition to the long-standing Cranfield paradigm popularized by information retrieval, this paper proposes an alternative "Manchester paradigm" closer to the Turing test, trivia games, and education.

Links:
- Research Talk [https://youtu.be/NZdKG3loB0M](https://youtu.be/NZdKG3loB0M)

Downloaded from [http://umiacs.umd.edu/~jbg/docs/2021_emnlp_paradigms.pdf](http://umiacs.umd.edu/~jbg/docs/2021_emnlp_paradigms.pdf)

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.
2 Serving Users: The Cranfield Paradigm

Let’s start with the Cranfield paradigm, named after Cranfield University in Bedfordshire. The Cranfield “experimental tradition founded by a librarian, working with card indexes, a half-century ago” spurred a revolution in information retrieval evaluation (Robertson, 2008).

In information retrieval, a “system locates information that is relevant to a user’s query” (Sanderson and Croft, 2012). The most natural IR evaluation is to ask users whether documents satisfy their information need. However, much like annotation in NLP, this is expensive and time-consuming, so Cleverdon (1967) proposes an alternative. Rather than have users interact with every potential system, build re-usable test collections and evaluate all systems by re-using the same collection. Although “obvious” to twenty-first century readers, the use of offline test collections for evaluation was controversial (Taube, 1965), and the approach is still debated (Saracevic, 2007).

Rather than putting users in front of every IR system, in the Cranfield paradigm, researchers perform experiments on test collections to compare the relative effectiveness of different retrieval approaches (Voorhees, 2002b).

Cranfield paradigm datasets approximate users’ searches, and the better your algorithm satisfies those queries, the better (Spärck Jones, 2001) the algorithm.

As IR systems conquered retrieving documents for short queries, researchers turned to finding short answers instead of whole documents (Voorhees, 2000b; Sanderson and Croft, 2012), which naturally lends itself to answering questions and “move[s] retrieval systems closer to information retrieval as opposed to document retrieval” (Voorhees and Tice, 2000). Under the Cranfield paradigm, a
good QA system should answer the questions users ask. What more could you want?

It is then unsurprising that Google and Microsoft adopted this setting for Natural Questions (Kwiatkowski et al., 2019) and MS MARCO (Nguyen et al., 2016). They found questions people asked online and answered them. A good QA system should automate that process (Chen and Yih, 2020).

3 Probing and Pushing Answerers: The Manchester Paradigm

In the Manchester paradigm, we create tasks and datasets whose questions push answerers to better understand the world and create evaluations that probe for human-like capabilities. Since we identify and name the paradigm, we give three justifications that highlight three distinct reasons people ask questions beyond information seeking: to teach, to compare, and to probe.

3.1 Why the Name Manchester?

For symmetry with the Cranfield paradigm, our proposed name is also an English city: Manchester. Because there are multiple aspects to the Manchester paradigm, we provide three connections between the city and question answering: in the nineteenth century, the city’s regiment used the mythical Sphinx as a symbol, it is the home to University Challenge, and it is where Alan Turing outlined the Turing Test. We discuss each of these reasons for the name Manchester paradigm (saving the best for last).

To Teach: The Sphinx, Manchester’s Standard

Manchester’s Regiment used the Sphinx as its symbol (Farmer, 1901). In Greek myth, the Sphinx asked everyone who entered the city a riddle (Renger, 2013). A Grammy-winning interpretation of the riddle states it as:

What starts out on four legs then goes round on two Then finishes on three before it’s through (Schickele, 1990)

And the answer is a human (crawling baby, walking, and then walking with a cane).

We have neither the space nor the desire to spoil Ïedipus’s journey, but by answering the question, the hero revealed not just his intelligence to his questioner but learned a tool to uncover his shrouded history. In a Whitman commencement, a classicist contrasted Google’s NQ with professors’ Sphinx-like riddles (emphasis added):

Nobody told Ïedipus who he was; he figured that out for himself. . . You can do a Google search to find out the capital of Arkansas (Little Rock?) but you can’t . . . find out who you are, what you’re good at, what makes you happy, what matters for your life . . . [The Sphinx’s riddle has a] hybrid nature: the curse of being forced to solve a riddle is also the gift of the ability to solve riddles. (Burgess, 2013)

This aspect of question answering also brings us to another Greek connection to asking questions: teaching through the Socratic method (Trepanier, 2017). Through asking the right questions, a teacher guides the answerer to understanding. While Cranfield questioners are seeking information from a more knowledgeable answerer, Manchester questioners often test less knowledgeable answerers. Similarly, perhaps by asking the right questions, the QA community can coax computers to understand more than they do now (Dunietz et al., 2020; Perez et al., 2020).

To Compare: Granada Studios

Another inspiration for this question answering approach is Manchester’s Granada Studios, creator of University Challenge (Taylor et al., 2012; Baber, 2015). This television programme juxtaposes two universities to see who is smarter.

Just like the Sphinx, the dapper host of this game show, Bamber Gasgoine, knows the answer. Thus it is not an information-seeking task a la the Cranfield paradigm. It, like the riddle of the Sphinx, is a test of those answering the questions.

It’s also a tried and true test of question answering researchers’ mettle, as when IBM Watson bested Ken Jennings on Jeopardy! (Ferrucci et al., 2010). While Cranfield focuses on users’ satisfaction, Manchester is at its heart an evaluation of the underlying capabilities of question answerers (either systems or humans): which is smarter, which is worthy? And as we discuss in Section 5, the Manchester paradigm is better suited to discriminating between answerers.

To Probe: The Turing Test

The final reason is a paper written by Alan Turing while at the University of Manchester. Rather than create a test for intelligence and be forced to face the substantial challenge of defining intelligence, Turing proposes an indistinguishability test (Turing, 1950). Building

1Likewise, we follow in Turing’s footsteps and sidestep the definition of intelligence. Like Brooks (1991), we would argue that the best way to test and refine any definition of intelligence would be to have an increasingly difficult sequence of QA challenges in the Manchester paradigm.
off what he imagined would be a fun Victorian-era party game called “the Imitation Game” (Bishop, 2010), a skilled interrogator would ask questions to either a machine or a computer. An intelligent computer should—at minimum—be able to make itself indistinguishable from a human. This competition, the Turing Test, has been called AI-complete (Yampolskiy, 2013) and when taken literally is the implicit basis for claims of “super-human AI” (Cuthbertson, 2018). Its ubiquity extends beyond computer science to popular culture. A variant in Blade Runner tests empathy—rather than intelligence—with probing questions (Joerden, 2012).

Likewise, for tests of intelligence in the Manchester paradigm, the Turing Test “represents what it is that AI must endevou[r] eventually to accomplish scientifically” (Harnad, 1992). Methodologically, the Manchester paradigm iteratively imagines tasks where machines should rival humans (Levesque, 2014), develops systems, and then determines if systems pass the test.

3.2 Examples

Questions derived from education (Clark, 2015), puzzles (Littman et al., 2002), and trivia competitions (Joshi et al., 2017) are in the Manchester camp (full categorisation in Appendix A). However, prominent Manchester paradigm questions were first composed for computers: the Winograd schema challenge (Levesque et al., 2011) and its successor the Winograd challenge (Sakaguchi et al., 2020). In this task, changing one word between two nearly identical binary questions also changes the answer. Should a machine fail such questions, it does not evince intelligence—at least not like humans.

While we set these paradigms in opposition to each other, we next discuss the swath of research that advances the goals of both.

4 What Cranfield and Manchester Share

Although these paradigms have different core goals, research advancing the goals of one often advances the goals of the other.

While there are differences between QA datasets across paradigms (Cambazoglu et al., 2020; Zeng et al., 2020; Dzendzik et al., 2021), these differences are overshadowed within a paradigm by dataset-specific quirks. Thus, a paradigm-agnostic blueprint for QA (Chen and Yih, 2020) is to combine sparse (Chen et al., 2017) or dense retrieval (Guu et al., 2020; Karpukhin et al., 2020) followed by span selection (Seo et al., 2017) or generation (Lewis et al., 2020). As a consequence, researchers indifferent to which questions are answered can improve representations and algorithms for both paradigms (although as interactions become richer, this may not be the case, as we discuss at the end of Section 6). The paradigms’ evaluations also overlap; they benefit from expert annotators (Gardner et al., 2020; Feng and Boyd-Graber, 2019), crowd annotators, and alternative evaluations like behavioural testing (Ribeiro et al., 2020). Similarly, both paradigms value robustness (Dalvi et al., 2004; Jia, 2020). Additionally, answering infrequently asked questions is important for search engines (Baeza-Yates et al., 2007), and building models that learn more from less qualifies as intelligent behaviour (Linzen, 2020). Creating systems robust to spelling mistakes (Wang and Pedersen, 2011) is a worthy goal. From the Cranfield perspective, systems hobbled by spelling mistakes lead to a poor user experience. On the other side, humans are impressively robust to poor spelling (Rayner et al., 2006), so from the Manchester perspective this form of robustness is also valuable. But this has its limits; in the next section, we argue why adversarial examples are more consistent with the Manchester paradigm.

5 Ignore the Distinction at your Peril

How Adversarial is too Much? Common ground has its limits. That there is not a dichotomy between these two approaches can sometimes mask the importance of distinguishing motivations. Other proposals for robustness postulate that models should be robust to input modifications users would not make (Feng et al., 2018), challenging yet unnatural adversarial questions that users are unlikely to ask (Jia and Liang, 2017; Wallace et al., 2019; Bartolo et al., 2020; Kiela et al., 2021), and testing a concept in multiple ways (Gardner et al., 2020; Kaushik et al., 2020). While solving these challenges may eventually improve Cranfield-motivated systems, in the short term solving these challenges does not directly contribute to improving the user experience: researchers who build overly artificial datasets are likely going to be ignored by the Cranfield-focused community.
Users are the Customers  Many future business dissertations will survey IBM Watson’s circuitous route from TREC system (Chu-Carroll et al., 2002) to Jeopardy! spectacle to embattled spin-off (Deutscher, 2021) despite “IBM [having] bragged to the media that Watson’s question-answering skills are good for more than annoying Alex Trebek.” (Jennings, 2011). One challenge may have been transitioning between paradigms. One aspect that made the transition more difficult was that the tour de force victory on Jeopardy! was firmly on the side of the Manchester paradigm, but to be a successful commercial application, it needed to make the shift to the Cranfield paradigm.

Similarly, SQuAD was written by people (Mechanical Turkers) who knew the answers...just like most of the questions in the Manchester paradigm. However, it did not follow the same principles of the Manchester paradigm, which led to the “shortcuts” that other investigators have discovered in the years since (Weissenborn et al., 2017). For example, priming made exploitable clues more frequent and Mechanical Turkers write each question as quickly as possible. Levesque (2014) anticipates this behaviour, specifically avoiding “cheap tricks” in their Manchester-paradigm Winograd challenge. In other Manchester paradigm questions, trivia question writers frequently take pride in well-crafted questions (Boyd-Graber and Börschinger, 2020).

Comparisons  One of the primary inspirations for the Manchester paradigm is competitions (e.g., University Challenge). Because these competitions are meant to determine who the smartest answerer is, they are remarkably efficient. The world accepted the judgement that Watson was smarter than Ken Jennings and Brad Rutter after 122 answers. Why not? These competitions are designed to discriminate between player abilities. In contrast, the dev and test sets of Cranfield-inspired datasets have thousands of questions, and even that may not be enough (Card et al., 2020; Rodriguez et al., 2021).

6 Call to Action

Our central plea is that researchers in QA and NLP more broadly should have a clear answer to the question: “Why are you working on this?” This is of particular importance as QA datasets proliferate (Rogers et al., 2021), and NLP practitioners “lost in dozens of recent datasets” want to know what datasets measure (Rogers and Rumshisky, 2020). While Gardner et al. (2019) offer a trenchant enumeration of QA uses, we think the onus of definition should fall on dataset creators, not on post-hoc analyses. Other than more explicitly naming two of the uses of QA after English University towns, our goal is to encourage researchers to recognize the tensions between these two uses and the opportunities created from recognizing the distinction.

Make what you Value Explicit  Each of these paradigms value different skills and embed these values in datasets and tasks. To make machine learning useful to society and adopt valuesensitive design (Dotan and Milli, 2020), developers of datasets should make their goals clear from the outset (Bender and Friedman, 2018; Gebu et al., 2018). In the Cranfield paradigm, aligning these evaluations with user satisfaction is essential (Spärck Jones, 2001). Industry is naturally financially motivated towards this goal, and they have the user data (Zhang et al., 2019)—only a fraction of which is published to protect privacy (Barbaro and Zeller, 2006). Still, strategic and thoughtful partnerships like the Cranfield-inspired TREC workshops are valuable; without TREC, it is estimated that “US Internet users would have spent up to 3.15 billion additional hours using web search engines between 1999 and 2009” (Tassey et al., 2010). One of the goals of the Manchester paradigm should be to identify the linguistic phenomena or ethnic and linguistic groups (Peskov et al., 2021) that are not well-served by Cranfield-focused data.

Thus, before you begin your question answering research, make it clear what your goal is: are you trying to build AGI or to serve users? That answer will then inform your evaluation methodology.

Academia’s Special Role  It is no coincidence that our paradigms are named after the homes of universities, and universities are where the Manchester paradigm will thrive. Thus, there is a strategic opportunity for academia and funding agencies to support Manchester-aligned work abjured by...
deep-pocketed industry. Lovingly crafted questions by trivia experts (Rodriguez et al., 2019; Boyd-Graber and Börschinger, 2020) and adversarial questions (Wallace et al., 2019; Bartolo et al., 2020; Kiela et al., 2021) are unlikely to change the way a smart assistant answers a question, but they might expose blind spots of QA systems or improve evaluation. Moreover, asking questions in public is not just entertaining; it can generate data (von Ahn and Dabbish, 2008) and help the public better understand the possibilities and limitations of AI (hsiung Hsu et al., 1995; Silver et al., 2016). Thus, those in the Manchester paradigm can game show-ify evaluations to make question answering more fun and illuminating.

7 Conclusion

We identify two core motivations for QA research over the past twenty years. We link one to the user-centered goals of the Cranfield paradigm and propose the Manchester paradigm to describe research working towards building human-like, intelligent QA systems. In at least the short-term, this distinction is important as it illuminates the goals of industry and academic stakeholders; ultimately, this makes it easier to ensure that both research agendas are valued. In the long term, we suspect that the best QA agents will benefit from the insights of user-oriented tasks and the longer-range efforts towards natural language understanding (Bender and Koller, 2020; Linzen, 2020).

Acknowledgements

For helpful comments and whose work was an inspiration for this paper, we thank Ellen Voorhees. For insightful discussions and ideas we thank Doug Oard, Shi Feng, and Alexander Hoyle. For feedback on prior versions of this paper, we thank members of the UMD CLIP lab, Robin Jia, Patrick Lewis, Joe Barrow, Eleftharia Briakou, Adina Williams, Douwe Kiela, and John P. Lalor. We thank our anonymous EMNLP reviewers and meta-reviewer for suggestions and comments. Boyd-Graber and Rodriguez’s work at UMD were supported by NSF Grant IIS-1822494. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsors.

References


Michael Barbaro and Tom Zeller. 2006. A face is exposed for aol searcher no. 4417749.


Peter Clark. 2015. Elementary school science and math tests as a driver for AI: take the aristo challenge! In Association for the Advancement of Artificial Intelligence, pages 4019–4021.


Cyril Cleverdon. 1967. The cranfield tests on index language devices. In Aslib proceedings. MCB UP Ltd.

Anthony Cuthbertson. 2018. Robots can now read better than humans, putting millions of jobs at risk. Newsweek.


Andrew Ng. 2015. Deep learning: What’s next.


Peter Schickele. 1990. Oedipus Tex and other choral calamities.


A Categorizing QA Datasets by Paradigm

To make our QA evaluation paradigms idea more concrete, we categorize fifty-six QA datasets as either primarily motivated by the Cranfield paradigm or the Manchester paradigm (Table 1). As is expected, the TREC QA tasks fall under the Cranfield paradigm while trivia-based datasets like Jeopardy! (SearchQA), Quizbowl, and TriviaQA fall under the Manchester paradigm. Many of the datasets that fall under the Manchester paradigm attempt to probe for “understanding” of some context; SQuAD for example probes for “understanding” of a context paragraph. Other datasets like ELI-5 are also clearly Cranfield since they are sourced specifically from questions that real users have asked. Although Table 1 likely does not enumerate all QA datasets, it nonetheless represents a extensive survey of the most prominent QA datasets. For more extensive QA surveys, see Cambazoglu et al. (2020) and Rogers et al. (2021) or a tutorial by Chen and Yih (2020).
Table 1: We categorize QA datasets by paradigm, domain area, and who authored them. For authorship, we show whether they are authored by non-crowdsourced humans (●), crowdsourcing (●), or are automatically generated (●). When humans and machines collaborate on questions, we indicate this and the directionality (e.g., SimpleQuestions generates knowledge base triples that crowdsource workers phrase as questions).