CausalQA: A Benchmark for Causal Question Answering

Alexander Bondarenko [⋆] **Magdalena Wolska** [†] **Stefan Heindorf** [‡] **Lukas Blübaum** [‡]
Axel-Cyrille Ngonga Ngomo [‡] **Benno Stein** [†] **Pavel Braslavski** [§,¶] **Matthias Hagen** [⋆] **Martin Potthast** [‖]

[⋆]Martin-Luther-Univeristät Halle-Wittenberg [†]Bauhaus-Universität Weimar
[‡]Paderborn University [§]Ural Federal University [¶]HSE University [‖]Leipzig University

Abstract

At least 5% of questions submitted to search engines ask about cause–effect relationships in some way. To support the development of tailored approaches that can answer such questions, we construct Webis-CausalQA-22, a benchmark corpus of 1.1 million causal questions with answers. We distinguish different types of causal questions using a novel typology derived from a data-driven, manual analysis of questions from ten large question answering (QA) datasets. Using high-precision lexical rules, we extract causal questions of each type from these datasets to create our corpus. As an initial baseline, the state-of-the-art QA model UnifiedQA achieves a ROUGE-L F_1 score of 0.48 on our new benchmark.

1 Introduction

The term "causality" usually refers to a directed relationship between events in which one is the cause of the occurrence of the other, called the effect. Many empirical studies begin with a research question about a causal relationship, ranging from "yes/no"-questions such as "Does the quality of education affect economic growth?" to open-ended questions such as "What causes depression?". But the general public also frequently asks causal questions. Figure 1 shows an example of the top Google, Bing, and Naver search result for the question "Can broccoli cause constipation?". While Bing directly answers the question in the affirmative, Google's featured snippet and Naver's first snippet claim that broccoli actually has the opposite effect.

With at most a few thousand question–answer pairs, existing causal question answering datasets are relatively small and include only one type of causal question, e.g., "yes/no"-questions (Hassanzadeh et al., 2019; Kayesh et al., 2020), "what-if"-questions (Tandon et al., 2019), "why"-questions (Verberne et al., 2006a, 2008, 2010; Lal et al., 2021), or multiple-choice questions (Gordon et al., 2012). The effectiveness of question answering (QA) systems on these benchmarks range from F_1 scores of 0.67 to 0.72. In contrast, QA systems have already performed better than humans for arbitrary questions. For instance, the F_1 score of the most effective system on the SQuAD benchmark is 0.93, while that of humans is only 0.89 (Rajpurkar et al., 2018).[1] Since neither SQuAD nor other large QA benchmarks explicitly label causal questions, the difference in effectiveness between causal and other questions remains unclear. But the inconsistent results of Bing compared to Google and Naver in Figure 1 suggest that more research is needed on answering causal questions.

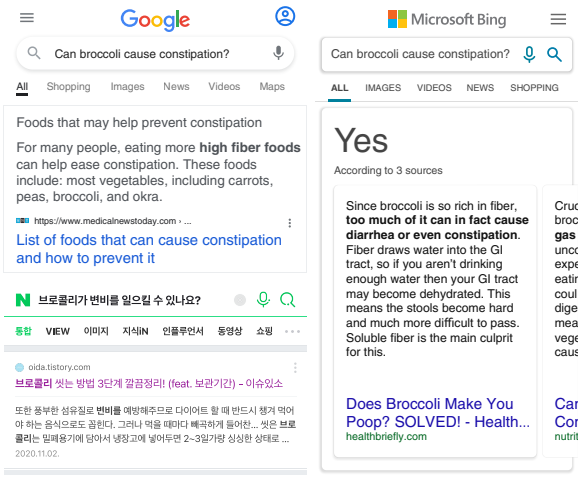

Figure 1: "Can broccoli cause constipation?" Google's and Naver's top results both disagree with that of Bing.

We take the first steps towards a more thorough investigation of causal question answering by creating the Webis-CausalQA-22 benchmark,[2] which consists of 1.1 million questions and answers about causal relationships.[3] The resource compiles causal questions from the ten QA datasets shown in Table 1. To identify the causal questions in these

[1] https://rajpurkar.github.io/SQuAD-explorer/
[2] Leaderboard: https://causalqa.webis.de
[3] Code and data: https://github.com/webis-de/COLING-22

Table 1: Characteristics of the question answering datasets used to create Webis-CausalQA-22. We removed questions without answer (respective datasets marked by *; in total, 25,841 causal questions without answers).

Dataset	Type		Size			Length (Words)		Reference
	Question source	Answer	Questions	Causal questions		Caus. qu.	Answ.	
PAQ	Generated with BART	Term(s)	64,875,601	769,606	(1.2%)	9.6	2.7	Lewis et al. (2021)
GooAQ	Google's autocomplete	Term, Passage	5,030,530	146,286	(2.9%)	7.3	44.3	Khashabi et al. (2021)
MS MARCO QnA*	Bing query log	Passage	1,010,916	25,569	(2.5%)	6.4	17.5	Nguyen et al. (2016)
Natural Questions*	Google query log	Passage	315,203	1,208	(0.4%)	9.8	10.8	Kwiatkowski et al. (2019)
ELI5*	Reddit questions	Passage	272,634	131,033	(48.0%)	32.5	99.0	Fan et al. (2019)
SearchQA	Human-written	Term(s)	216,136	780	(0.4%)	16.8	1.8	Dunn et al. (2017)
SQuAD v.2.0*	Human-written	Term(s)	142,192	3,209	(2.3%)	10.5	6.2	Rajpurkar et al. (2016)
NewsQA*	Human-written	Term(s)	119,633	652	(0.5%)	7.2	6.1	Trischler et al. (2017)
HotpotQA	Human-written	Term(s), Passages	112,662	390	(0.4%)	21.8	3.8	Yang et al. (2018)
TriviaQA	Human-written	Term(s)	109,767	703	(0.6%)	19.4	3.1	Joshi et al. (2017)
Webis-CausalQA-22	Mixed	Mixed	72,205,274	1,079,436	(1.5%)	12.0	22.5	This paper

datasets, we manually analyzed samples and developed a two-dimensional typology of causal questions based on their semantic properties and pragmatic interpretation (Section 3). Using a set of manually created lexical rules, we extract causal questions with 80% recall at near-perfect precision (Section 4). When applied to a large sample of a query log from a commercial search engine, we also find that at least 5% of submitted queries are causal, highlighting the need for tailored technologies. As an initial baseline, we evaluate the UnifiedQA model (Khashabi et al., 2020) fine-tuned on our resource (Section 5). It achieves an average ROUGE-L F_1 score of 0.48 across datasets.

2 Related Work

We review the literature in four areas: prior typologies of causal questions, causal QA, as well as generic QA datasets and QA systems.

2.1 Causal Question Typologies

In the QA literature, causal questions are usually considered in terms of their lexical surface form and their answer type (i.e., the content of the answer). Most of the existing causal question typologies only deal with questions clearly identifiable by the question word "why". Somewhat consequently, early open-domain QA research only had a single type covering all "why"-questions (Hovy et al., 2000; Moldovan et al., 2000, 2003) before Verberne et al. (2006b) subcategorized them based on the answer type as cause (no deliberate human intention involved), motivation (human intention involved), circumstance (strict condition for the resulting event), or generic purpose (physical function of an object). For Webclopedia, Verberne et al. (2007) suggested five types: motivation, physical explanation, non-physical explanation, etymology, and nonsense. Later, Breja and Jain (2017) proposed another, rather reasoning-based, typology of causal questions: informational / factual (reasoning about a fact), historical (reasoning about the past), situational (reasons for an event at a particular time), and opinion (personal reasons).

Interestingly, all these typologies lack abstraction and do not capture general properties of causal relations. For instance, physical, non-physical, and etymology can be seen as subtypes of a class "causal explanation" that specify the nature of the explanation. The typologies also operate at different granularities, which makes comparisons difficult. For instance, Verberne et al. (2007) address specific properties of causes (physical, linguistic), whereas Breja and Jain (2017) focus on the strength of the evidence (fact vs. opinion).

In contrast, an objective, data-neutral approach to categorizing questions in general had been proposed by Lehnert (1977), including some causal types dependent on the structure of the causal dependencies. Our typology builds on Lehnert's, and we derive subtypes of causal questions in a systematic way along with their semantic and pragmatic characteristics: analytically at the semantic level and in a data-informed fashion at the pragmatic level. Moreover, our approach is not limited to causal "why"-questions as in most of the prior work, but characterizes the type of causal questions independent of their surface form.

2.2 Causal Question Answering

The related work on causal QA is rather limited. Most datasets for causal QA focus on "why"-

questions and are relatively small (Gordon et al., 2012; Hassanzadeh et al., 2019; Verberne et al., 2006a, 2008, 2010; Tandon et al., 2019; Kayesh et al., 2020; Lal et al., 2021). Usually, QA systems only achieve F_1 scores of around 0.7 on these datasets—worse than the effectiveness observed for many other question types. For instance, Ishida et al. (2018) and Iida et al. (2019) retrieve "compact" answers for "why"-questions from a web corpus using a pointer-generator network (See et al., 2017). Kayesh et al. (2019) address causal "yes/no"-questions by transfer learning, while Hassanzadeh et al. (2019) use large-scale text mining. Finally, Heindorf et al. (2020) suggest to use CauseNet, a large knowledge graph with more than 11 million cause–effect relationships extracted from ClueWeb12 web pages and Wikipedia articles. With Webis-CausalQA-22, we create a larger dataset to enable training and testing causal QA approaches on a dedicated broader benchmark.

2.3 Question Answering Datasets

Current QA research is characterized by the growing sizes of datasets (see Table 1) to improve neural QA models, and by a diversification across domains and question types (e.g., HotpotQA specifically includes comparative questions) and languages (e.g., TyDi QA features eleven languages). QA systems have meanwhile outperformed humans on Rajpurkar et al.'s (2018) SQuAD benchmark for reading comprehension. Hence, new task-specific smaller benchmarks such as CommonsenseQA (14,000 "yes/no"-questions by Talmor et al. (2019, 2021)) for common sense reasoning have been published as new challenges. On CommonsenseQA v. 2.0, for instance, Lourie et al.'s (2021) T5-based UNICORN model achieves an accuracy of 0.7, but this is still below the 0.94 of humans (Talmor et al., 2021). Out of the many available open-domain QA datasets, we selected those that are well-known enough to be mentioned in surveys (e.g., Cambazoglu et al. (2020)), contain lexically diverse question types, and have more than 100,000 QA pairs (cf. Table 1 for the selected datasets and their characteristics).

Artificial datasets. With 65 million QA pairs, PAQ (Lewis et al., 2021) is the largest of the selected datasets. Its questions were generated using the BART-base model (Lewis et al., 2020) fine-tuned on the questions, answers, and context passages from Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and SQuAD (Rajpurkar et al., 2016). Fine-tuning on human questions ensures some naturalness, but the answers were automatically extracted from Wikipedia. Among our selected datasets, PAQ is the only automatically generated one. We include it since the generation evaluation by Lewis et al. shows the questions to be plausible and since more than 700,000 causal questions are contained.

User-generated datasets. GooAQ (Khashabi et al., 2021), MS MARCO QnA (Nguyen et al., 2016), Natural Questions (Kwiatkowski et al., 2019), ELI5 (Fan et al., 2019), SearchQA (Dunn et al., 2017), and TriviaQA (Joshi et al., 2017) contain real-world questions submitted to search engines or posted on web fora. The GooAQ dataset contains about five million QA pairs with questions collected from Google's query auto-completion when prompted with a given question word. The answers were extracted from Google's featured snippets shown as direct answers on top of the search results. The MS MARCO QnA corpus contains about one million questions that were sampled from Bing's query logs, with long answers (text passages) extracted from web documents retrieved by Bing, and short answers (terms) written manually by crowdworkers. Similarly, the Natural Questions dataset contains more than 300,000 queries sampled from Google's search logs. Long answers and short answers were manually selected by crowdworkers from Wikipedia articles.

The about 270,000 ELI5 questions were collected from Reddit's subreddit "Explain Like I'm Five (ELI5)" where users provide simple answers to posted questions. Only answers (text passages) with at least two more up-votes than down-votes were used. The more than 215,000 questions in SearchQA and their short answers (terms) stem from Jeopardy!, while context passages were obtained by querying Google and collecting at least 40 result snippets. The more than 100,000 QA pairs in TriviaQA were scraped from various trivia and quiz websites. Each QA pair is complemented with context passages in the form of web documents from Bing's search results or from Wikipedia.

Crowdsourced datasets. The SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), and HotpotQA (Yang et al., 2018) datasets were exclusively constructed using crowdsourcing. SQuAD version 2.0 contains about 140,000 QA pairs written by crowdworkers who were shown paragraphs

from Wikipedia and tasked to compose up to five questions and answers about them. The about 120,000 QA pairs in NewsQA were similarly crowdsourced using CNN news articles' headlines and their summaries, but different crowdworkers wrote the questions and the answers. Lastly, HotpotQA contains about 113,000 entries with questions, answers, and supporting facts written by crowdworkers based on Wikipedia paragraphs. Designed for multi-hop QA, these questions require a system to "hop" over several supporting facts (mostly sentences) from different text passages to arrive at a short answer.

2.4 Question Answering Systems

Early question answering systems such as Baseball (Green et al., 1961) used dictionaries of attribute–value pairs to answer questions, usually in narrow domains. Recent, more sophisticated QA systems can be divided into systems based on textual data and systems based on knowledge graphs.

Text-based systems, like UnifiedQA (Khashabi et al., 2020) that we employ as a first baseline for our new benchmark, mainly use language models. Their input may just be a question but often also is a question with context like some text passage or even the whole Wikipedia (Chen et al., 2017). The actual answering process ranges from binary classification (answer selection) over span extraction (identifying answer boundaries within a text) to abstractive text summarization and generation.

Knowledge base question answering (KBQA) systems operate on graphs with a single or up to thousands of edge types (e.g., DBpedia by Auer et al. (2007)). Typically, they use manually designed templates of graph patterns to detect answers (Zheng et al., 2018; Vollmers et al., 2021), use knowledge graph embeddings (Sharp et al., 2016; Huang et al., 2019; Saxena et al., 2020), or train neural networks on knowledge graphs (Chakraborty et al., 2021). Questions are often divided by their answer type being a single graph relation (Mohammed et al., 2018), a path with multiple hops (Saxena et al., 2020), or complex answers requiring reasoning (e.g., combining information from multiple paths; Lu et al. (2019); Mitra and Baral (2016); Asai et al. (2020)).

3 A Typology of Causal Questions

While various types of causality-related questions have been previously addressed in automated question answering, there has been no attempt so far to systematize "questions about causality" as a class in the QA community. Computational processing of causal structures, however, dates back to the 1970s and the early AI research on causal dependencies between events in the context of story comprehension. Notably, Lehnert (1977) developed a computational model of question answering based on a theory of "conceptual information processing". Their QUALM system was capable of answering 13 types of questions about stories—9 types being related to causal relationships.

Following Lehnert, we define questions about causal relationships in terms of causal chains (Schank, 1975) and integrate Lehnert's causal categories into a more specific typology of causal questions. While Lehnert's definitions and categories are motivated by and directly linked to processing strategies in a story comprehension system, our typology is more generic and motivated by the semantic and pragmatic properties of causal questions. At the semantic level, we group causal question types in terms of which component of a causal chain a question addresses. Our type set combines Lehnert's causality-related categories and Verberne et al.'s categories of "why"-questions (Verberne et al., 2006b, 2007) as subtypes. At the pragmatic level, we group question types in terms of the assumed purpose of inquiry or the so-called *intent* of a question. We arrived at the pragmatic categories in a data-driven fashion by analyzing 1,000 questions (100 sampled from each of the 10 selected QA datasets; cf. Table 1). In the following sections, we first define the category *causal question* and then present the semantic and pragmatic dimensions of our typology.

3.1 The Causal Question Category

We define the category *causal question* by referring to knowledge resources required in providing an answer, specifically, to inference based on causal chains (Schank, 1975). A *causal chain* is a sequence of alternating events (or statestions) linked by relationships expressing causal dependencies between them: an event can *enable*, *result in*, be the *reason of*, *cause*, or *lead to* another event. A question is a *causal question* if answering it requires (1) identifying causal chains, (2) inference on those chains, and (3) verbalizing the causal relations involved when answering it. By this definition the question "Why is there something rather than noth-

Table 2: (a) The semantic and (b) the pragmatic dimensions of causal questions; the set of subtypes in (b) is not exhaustive, but serves to show that the top-level categories are well-motivated—considering that coherent subtypes can be identified—and to illustrate the range of domains of the requests. (c) Rules to classify causal questions in the labeled sample of 1,000 questions. Reported: precision and recall for the class of causal questions and number of matches in Webis-CausalQA-22. For the rules not present in the initial random sample, we sampled 50 random questions afterwards, manually labeled them, and calculated a precision (numbers are given in gray).

(a)

Category	Examples
Questions about an antecedent	
Cause	Why does a mosquito bite itch?
Goal	Why did Jean Valjean take care of Cosette?
Purpose	Why do gaming chairs have a race car design?
Enablement	How can FIFA be so blatantly corrupt?
Questions about a consequent	
Result	What does increasing water vapor lead to?
Questions about the causal chain	
Verification	Would hydrophobic coating affect swimming?

(b)

Intent	Examples
Solution seeking	
Problem solving	
Practical problems	Why can't I log in into Facebook?
Medical problems	Can broccoli cause constipation?
Problem prevention	
Medical problems	What to do to prevent cancer?
Societal problems	What to do to prevent global warming?
Coping with problems	
Mental coping	Why do you think about the people who are gone?
Anger	Why doesn't a director fire a stupid employee?
Knowledge seeking	
Physical world	Why do chemical reactions depend on pH?
Politics / history	How did World War II start?
Language	Why is a notebook called "notebook"?
Trivia / fun facts	What happens if you scan a mirror?
Opinion seeking	
Social issues	Why do men cheat on their wives?
Entertainment	Why is Messi not playing on the team?
Rational future outcomes	What will happen if Trump wins another election?
Irrational future outcomes	What will happen if one dreams of pregnancy?

(c)

Measure	Lexical rules							
	R1	R2	R3	R4	R5	R6	R7	R1–7
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.59	0.11	0.07	0.02	0.01	–	–	0.80
Matches	505K	313K	132K	131K	10K	15K	4K	1.1M

ing?" can be interpreted as causal and eliciting a physical cause for existence, whereas the question "What is your name?" will not be considered causal even though a causal chain leading to a person being given a name may be identified; the answer "My name is Mary" does not verbalize the causal chain and an answer like "My mother named me Mary", while it may be considered as related to the question, provides irrelevant information under standard assumptions about responsiveness.

3.2 The Semantic Dimension

Our three top-level semantic categories for causal questions reflect the question's target with respect to the structure of causal chains: *questions about an antecedent* ask about events, actions, or states that in a (maybe just hypothetical) causal chain precede the ones mentioned in the question, *questions about a consequent* ask about events, actions, or states that follow the ones mentioned in the question, and *questions about the chain* ask about some property of the causal chain itself. Each of these three categories has further more specific subtypes; selected subtypes with example questions are given in Table 2a. Note that the list of subtypes is not meant to be exhaustive: we show only those types that we actually encountered in the literature or in our annotated datasets.

Cause questions ask about a general cause due to which the consequent holds; the causal dependency may be of any type: physical cause, social, psychological, etc. *Goal* questions ask about intentional motives behind an action, be it general future goals or inner motivations, whereas *purpose* questions ask about a generic purpose of the consequent, and *enablement* questions ask about the circumstances that enable / enabled the consequent. *Result* questions ask about the general effect of the antecedent, and *verification* questions ask whether a causal chain between events exists.

In this typology, Lehnert's *disjunction* is subsumed under the more general *verification* category (properties of the verified proposition possibly marked as attributes) and *expectational* is an attribute of *cause*, since the only difference between Lehnert's *cause* and *expectational* categories is that in the case of the latter, the consequent act presumably did not occur. Verberne et al.'s *motivation*, essentially a combination of Lehnert's *goal* and *circumstance*, is our *enablement* category with possibly Charniak's (1975) additional attributes.

Note that answering procedural questions (e.g., "how to . . . ") also often involves inferences based on somewhat "causal" chains. However, procedural questions usually reflect a non-causal underlying information need in the sense that they ask about

the sequential nature of a chain but not about the causal relations. Such questions can rather be considered *manner questions*, as also suggested by Hovy et al. (2002), so that we do not include them in our typology of causal questions.

3.3 The Pragmatic Dimension

At the pragmatic level, we model the inquirer's assumed motive for asking, i.e., their "visceral need" in Taylor's (1962) terminology or the "query intent" in Broder's (2002). We link the causal questions' intents to the pragmatic function of the inquiries—their "illocutionary force" (Austin, 1962). Much as recognizing the underlying function of a question affects a listener's response strategy, also in the case of web search, being able to identify a query's underlying speech act can guide the choice of what resources (e.g., what document subset) to use in a search for answers.

Our analysis of the 1,000 question sample dataset revealed three core categories of intent in causal questions: *solution seeking*, *knowledge seeking*, and *opinion seeking*. These intents, in turn, can be interpreted in terms of their illocutionary force as indirect *requests for help* (some of the questions under *solution* and *opinion*) or as genuine *requests for information* (*solution* and *knowledge*). *Solution seeking* and non-trivia/-trivial *knowledge seeking* calls for search in authoritative knowledge sources, whereas *opinion seeking* calls for search, for instance, in discussion fora or social media. Moreover, recognizing a *request for help* (falls into *coping with problems* in our typology) in a question might justify additional content in the generated response, such as advice where to seek help in case of a medical question. Subtypes of the three intent categories are exemplified in Table 2b. Again, the presented subtypes are not meant to be exhaustive, but rather to show that the top-level categories are well-motivated and to illustrate the range of possible information needs in causal questions.

3.4 Causal Questions in Web Search

Finally, to gain some insights into causal questions that people actually submit to search engines, we briefly analyze a dataset of all question-like queries submitted to Yandex in 2012; dataset created by Völske et al. (2015) and also used by Bondarenko et al. (2020). The question-like queries were extracted from the complete 2012 Yandex log by matching any of 58 hand-crafted syntactic question indicators (e.g., queries starting with "how", "what", "where", etc.). The final set contains about 1.5 billion question-like log entries with about 730 million unique questions. Applying translated versions of the seven lexical rules we use for our benchmark corpus construction (cf. Table 3), about 81.7 million causal questions are mined from the log (about 5% of the 1.5 billion question-like log entries). The most frequent causal questions are "why"-questions (50 million; frequent example: "Why can't I log in into VKontakte?") followed by "what to do if"-questions (13.1 million) and "what happens if"-questions (11 million). Interestingly, from manual spot checks of 1,000 mined "what happens if"-questions, it seems that 90% of them are about dream interpretation (e.g., "What will happen, if one dreams of pregnancy?"). This category of somewhat *fictitious* causality, raises the interesting question about how search engines or QA systems should deal with respective information needs. However, somewhat unsurprisingly, such examples are not contained in current standard QA datasets. Another manual inspection of a sample of 1,000 questions explicitly asking about causes or effects shows that most of them target causes of medical conditions or effects on health.

4 The Webis-CausalQA-22 Corpus

In this section, we describe how we extract causal questions from the ten QA datasets in Table 1 and briefly analyze or resulting new benchmark corpus.

4.1 Corpus Construction

Table 1 gives an overview of the QA datasets from which we extract causal questions to construct the Webis-CausalQA-22 benchmark. The datasets fulfill three selection criteria: (1) they contain lexically diverse questions, (2) they are well-known in the research community, and (3) they are large.

We investigate causal questions in two steps: based on prior work and based on a manual analysis of 1,000 questions randomly sampled from the QA datasets (100 from each). We asked two annotators to label whether a given question is causal or not, considering a question to be causal if the answer may only be provided as a result of causal reasoning involving entities from the question. To discover new patterns beyond more "obvious" ones like "What are the effects of X?" or "What causes Y?", we did not provide examples, but specified that the question may be asking about explicit or implicit causal relationships. They achieved an

ID [Regular Expression] / Example
R1 `[why]` Why does mosquito bite itch?
R2 `[cause(s)?]` What causes broken blood vessels?
R3 `[how come|how did]` How did the constellation Bootes get its name?
R4 `[effect(s)?|affect(s)?]` What was the effect of the silk road on religions?
R5 `[lead(s)? to]` What does increasing water vapor lead to?
R6 `[what (will|might)? happen(s)?]∧[if|when]` What happens if we drink very hot water?
R7 `[what (to do|should be done)]∧[if|to|when]` What to do if my Xbox won't connect to the Wi-Fi?

Table 3: Lexical rules used to match causal questions in a regular expression syntax. E.g., a question matching R6 must contain 'what happens' or 'what will happen' or 'what might happen' and 'if' or 'when'.

inter-annotator agreement of Cohen's $\kappa = 0.54$ (moderate agreement). Coding differences were reconciled in a discussion with a third annotator and a total of 86 questions labeled as causal.

Based on the causal questions from our sample and based on existing question typologies (Lehnert, 1977; Graesser and Person, 1994; Graesser et al., 2008; McClure et al., 2001; Gelman, 2011; Gelman and Imbens, 2013), we hand-crafted the seven lexical rules to identify causal questions (cf. Table 3). Rules R1–R5 achieve a precision of 1.0 on our labeled sample (cf. Table 2c), while no instances matched Rules R6 and R7 (derived from prior work). We thus randomly sampled 50 questions from the QA datasets using these rules and manually checked that their precision also is 1.0.

We run these seven high-precision rules on the ten standard QA datasets and extract a total of about 1.1 million causal questions that, together with their answers and context passages (if available), form the Webis-CausalQA-22 benchmark corpus.

4.2 Corpus Analysis

Table 2c shows how many causal questions have been extracted by each of the seven lexical rules. About half of the causal questions are open-ended "why"-questions (e.g., "Why does a mosquito bite itch?"). Questions about causes (e.g., "What causes broken blood vessels?") constitute another 28% of our corpus. Interestingly, the least frequent ones are "what to do if"-questions (e.g., "What to do if my Xbox won't connect to the Wi-Fi?") that at less than 1% are by far less common than their 11% in real web search questions (cf. Section 3.4).

The context available for the question–answer pairs in our Webis-CausalQA-22 corpus depends on the source dataset and varies from Wikipedia passages (e.g., PAQ, Natural Questions, SQuAD) to search engine snippets (e.g., SearchQA) or passages from web documents (e.g., MS MARCO QnA). Also the average question and answer lengths vary widely per extraction source. While, on average, a question contains 12 words (cf. Table 1), the questions from MS MARCO QnA, for instance, are much shorter (6.4 words, Bing search) and questions from ELI5 are much longer (32.5 words, extracted from Reddit). Similarly, on average, an answer has 23 words but the answers from SearchQA are way shorter (1.8 words, human-written answers for Jeopardy! clues) while the answers from ELI5, again, are much longer (99 words, human-written answers with explanations). Besides the causal nature of the questions, also this diversity of questions and answers in our corpus poses a challenge to (causal) QA systems.

5 Evaluation

To establish a first baseline effectiveness for causal question answering on the Webis-CausalQA-22 benchmark, we report the results achieved by the state-of-the-art UnifiedQA model Khashabi et al. (2020, 2022). UnifiedQA is a text-based question answering model that has been reported by Khashabi et al. to perform well on 32 QA datasets, including SQuAD v. 2.0, where it achieved a bag-of-word-based F_1 score of 0.90. We experiment with Version 2 of the model, Checkpoint 1363200, using (1) the base model, and (2) a version fine-tuned on Webis-CausalQA-22 using the hyperparameters of Khashabi et al. (2022).[4] In a pilot study, we attempted to fine-tune a joint model on all datasets but found fine-tuning per source dataset to yield better results. Moreover, we experiment with the causal questions extracted from the original train/dev splits proposed by the authors as well as a random 90/10 train-test split of our own. The reason for the latter is that, for some datasets, by chance, only few causal questions are part of the original train/dev splits (compare the number of

[4] All experiments were conducted on an NVIDIA A100 GPU. Fine-tuning: 60K steps in general, or 6K steps to avoid overfitting on datasets containing less than 50K causal questions; AdamW optimizer (Loshchilov and Hutter, 2019); learning rate $5e^{-5}$; batch size 2.

Table 4: Effectiveness of the UnifiedQA model on causal question answering on the Webis-CausalQA-22 corpus: (a) the base model (Version 2) and a fine-tuned version on the original train/dev splits per dataset if availalbe; (b) a fine-tuned version on a random 90/10 train/test split. N: number of questions used for evaluation, P: precision, R: recall, F_1: F_1 score, EM: exact match. The star (*) indicates datasets usually evaluated using ROUGE-L.

Dataset	**Original train/dev split**											**Random 90/10 split**					
	N	Base model					Fine-tuned model					N	Fine-tuned model				
		ROUGE-L			Traditional		ROUGE-L			Traditional			ROUGE-L			Traditional	
		P	R	F_1	EM	F_1	P	R	F_1	EM	F_1		P	R	F_1	EM	F_1
PAQ	76,961	0.79	0.85	0.80	0.69	0.80	0.95	0.95	0.94	0.91	0.94	76,961	0.95	0.95	0.94	0.91	0.94
GooAQ*	33	0.29	0.04	0.06	0.00	0.07	0.14	0.11	0.12	0.00	0.15	14,629	0.17	0.15	0.15	0.00	0.19
MS MARCO QnA*	2,558	0.44	0.19	0.23	0.05	0.24	0.49	0.40	0.39	0.10	0.41	2,557	0.45	0.42	0.39	0.13	0.40
Natural Questions	71	0.14	0.05	0.06	0.01	0.07	0.34	0.37	0.33	0.18	0.34	121	0.37	0.34	0.32	0.16	0.33
ELI5*	13,104	0.26	0.04	0.06	0.00	0.08	0.16	0.09	0.10	0.00	0.12	13,104	0.16	0.09	0.10	0.00	0.12
SearchQA	117	0.20	0.22	0.20	0.15	0.20	0.63	0.64	0.62	0.53	0.62	78	0.55	0.54	0.54	0.47	0.54
SQuAD v.2.0	252	0.79	0.81	0.78	0.63	0.78	0.84	0.84	0.83	0.66	0.83	321	0.96	0.96	0.95	0.93	0.95
NewsQA	29	0.57	0.55	0.53	0.31	0.53	0.65	0.56	0.58	0.45	0.58	66	0.76	0.76	0.73	0.58	0.73
HotpotQA	35	0.49	0.39	0.40	0.14	0.40	0.60	0.55	0.53	0.26	0.54	39	0.73	0.73	0.73	0.67	0.72
TriviaQA	66	0.37	0.35	0.34	0.26	0.35	0.43	0.41	0.40	0.27	0.40	71	0.44	0.43	0.42	0.28	0.42
Macro-averaged	9,323	0.43	0.35	0.35	0.23	0.35	0.52	0.49	0.48	0.34	0.49	10,795	0.55	0.54	0.53	0.41	0.53
Micro-averaged	65,447	0.70	0.72	0.68	0.58	0.68	0.82	0.81	0.81	0.75	0.81	58,505	0.73	0.72	0.72	0.65	0.73

causal questions reported in Table 1 to the left Subcolumn N in Table 4. The original test sets are often not publicly available, but only indirectly via run submission to a leaderboard.

Effectiveness is measured using the ROUGE-L scores precision, recall, and F_1 (Lin, 2004), as well as the traditional exact match (EM) and F_1 measures. The ROUGE-L measures are based on the longest common subsequence between a predicted answer and a ground truth answer, whereas EM requires the order of all tokens to match and the traditional F_1 measure is based on an order-invariant bag-of-words representation. If a question has more than one ground truth answer, the maximum score per measure and question is taken. Effectiveness is measured both per constituent dataset of Webis-CausalQA-22, and averaged using micro- and macro-averaging across datasets.

Table 4 shows the effectiveness scores achieved by UnifiedQA. The columns "Original train/dev split" shows the effectiveness on the causal questions that we have identified in the original dev split using our lexical rules, yielding the number of causal question–answer pairs indicated in Subcolumn N.[5] We observe that UnifiedQA is most effective on PAQ across all measures, perhaps due to the large number of questions–answer pairs available and/or the fact that the models underlying PAQ and UnifiedQA have both been trained (among others, respectively) on SQuAD. For GooAQ and ELI5, the effectiveness is lowest, perhaps due to the lack of context information in these datasets. Fine-tuning UnifiedQA on the respective datasets increases its effectiveness in terms of ROUGE-L F_1 across the board. Overall, the scores of the fine-tuned models are between 0.12 and 0.62 with the exception of PAQ (0.94) and SQuAD (0.83). This generally indicates plenty of room for improvements in causal QA.

The columns "Random 90/10 split" reports the corresponding effectiveness scores of UnifiedQA for the fine-tuned model version, where fine-tuning was repeated on the different training set. Comparing the ROUGE-L F_1 scores to the fine-tuned model on the original split, we observe the largest differences for the datasets HotpotQA (from 0.53 in the original split to 0.73 in the new one), NewsQA (from 0.58 to 0.73), SQuAD (from 0.83 to 0.95), SearchQA (from 0.62 to 0.54), and GooAG (from 0.12 to 0.15). The effectiveness on HotpotQA increases because the original split used more difficult questions for the dev set than for training (Yang et al., 2018). The effectiveness on the new splits of SQuAD v. 2.0 and NewsQA increases because the UnifiedQA base model was trained on both datasets causing a leakage of training data. The effectiveness on SearchQA decreases potentially due to overfitting to the training set, or a particularly easy dev set in the original dataset by chance as the original dataset was split by time (different years for dev and test sets than for training). The effectiveness on GooAQ increases slightly because the original train/dev sets were explicitly made dissimilar by avoiding word overlaps while

[5] For PAQ and ELI5, no dedicated dev sets are available and we performed a random 90/10 split.

our random split does not. Moreover, with the new split, GooAQ is evaluated on many more questions because the original dev set contained far fewer causal questions than 10% of the whole dataset.

Overall, when comparing macro-averages across datasets, we find that fine-tuning improves the macro-averaged ROUGE-L F_1 scores from 0.35 to 0.48 on the original train/dev split, and to 0.53 on the random 90/10 split. Micro-averaging generally results in higher scores when compared to macro-averaging due to imbalanced distribution of question–answer pairs across datasets, where PAQ has the largest influence. Interestingly, when comparing the macro-averaged ROUGE-L F_1 scores of the original train/dev split with the random one, and the corresponding micro-averaged ones, the micro-averaged ones decrease from 0.81 to 0.72, while the macro-averaged ones increase as mentioned above. This is mainly caused by GooAQ having a much higher weight overall due to contributing more than 14,500 question–answer pairs, the second-largest amount following PAQ, compared to only 33 in the original train/dev split.

It is a matter of debate, which of the two splits and which of the two averages are to be preferred as a baseline. At present, we recommend using the original train/dev split (especially, if a model was trained on one of our corpus's constituent datasets, like UnifiedQA), and then the macro-averaged ROUGE-L F_1 score. In case of UnifiedQA, this score is 0.48 for the model version fine-tuned on each constituent dataset individually.

6 Conclusion

We constructed Webis-CausalQA-22, the first large benchmark dataset of 1.1 million causal question–answer pairs, which serves to advance research in causal question answering. To ensure diversity of questions, we extracted them using seven hand-crafted high-precision lexical rules to capture as many subtypes of causal questions as possible. These rules were derived from a new typology of causal questions, which in turn is based on relevant related work on question typologies. A manual analysis of a sample of questions was used to characterize causal questions in terms of two dimensions: (1) their semantic properties, i.e., according to which element of the causal structure the question is asked (antecedent, consequent, or the causal chain) and (2) their pragmatic interpretation, i.e., the underlying intention or assumed information need of the questioner (e.g., prevention of medical problems). Furthermore, a subsequent analysis of the causal questions contained in a search engine log showed that a significant proportion of 5% of question queries are causal. Finally, we evaluated the state-of-the-art model UnifiedQA on our corpus as an initial baseline for causal question answering.

Causal questions represent a hitherto poorly considered challenge for both search engines in general and QA systems in particular. In this respect, our typology serves as a guide for the development of new technologies: The semantic dimension is relevant for understanding queries, while the pragmatic dimension may guide search engines and QA systems in finding and presenting answers. In addition, linking current text-based models with algorithms for causal inference is a promising direction to answer more complex questions for which answers cannot be found directly on the web. CauseNet may also prove useful here, as the graph of cause–effect relationships already makes such connections. However, to maximize user confidence in an information system's answers to causal questions, all causal claims must be supported by evidence (e.g., in the form of scientific studies).

Acknowledgements

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). This work has also been partially supported by the German Federal Ministry of Education and Research (BMBF) within the project EML4U under the grant no 01IS19080B and within the project COLIDE under the grant no 01I521005D as well as by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the project RAKI under the grant no 01MD19012B. The authors acknowledge the Paderborn Center for Parallel Computing (PC^2) for providing computing resources. Pavel Braslavski acknowledges funding from the Ministry of Science and Higher Education of the Russian Federation (project 075-02-2022-877). We thank Yandex for granting access to the data.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learn-

ing to retrieve reasoning paths over Wikipedia graph for question answering. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC 2007 + ASWC 2007)*, pages 722–735.

John Langshaw Austin. 1962. *How to do Things with Words*. Oxford University Press, Oxford, UK.

Alexander Bondarenko, Pavel Braslavski, Michael Völske, Rami Aly, Maik Fröbe, Alexander Panchenko, Chris Biemann, Benno Stein, and Matthias Hagen. 2020. Comparative web search questions. In *Proceedings of the Thirteenth ACM International Conference on Web Search and Data Mining (WSD 2020), Houston, TX, USA, February 3–7, 2020*, pages 52–60.

Manvi Breja and Sanjay Kumar Jain. 2017. Why-type question classification in question answering system. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation*, pages 149–153.

Andrei Z. Broder. 2002. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

B. Barla Cambazoglu, Mark Sanderson, Falk Scholer, and W. Bruce Croft. 2020. A review of public datasets in question answering research. *SIGIR Forum*, 54(2):5:1–5:23.

Nilesh Chakraborty, Denis Lukovnikov, Gaurav Maheshwari, Priyansh Trivedi, Jens Lehmann, and Asja Fischer. 2021. Introduction to neural network-based question answering over knowledge graphs. *WIREs Data Mining Knowl. Discov.*, 11(3).

Eugene Charniak. 1975. A partial taxonomy of knowledge about actions. In *Advance Papers of the Fourth International Joint Conference on Artificial Intelligence (IJCAI 1975)*, pages 91–98.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1870–1879.

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*, abs/1704.05179.

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

Andrew Gelman. 2011. Causality and statistical learning. *American Journal of Sociology*, 117(3):955–966.

Andrew Gelman and Guido Imbens. 2013. Why ask why? Forward causal inference and reverse causal questions. NBER Working Papers 19614, National Bureau of Economic Research, Inc.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 Task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2012)*, pages 394–398.

Art Graesser, Vasile Rus, and Zhiqiang Cai. 2008. Question classification schemes. In *Proceedings of the Workshop on Question Generation*, pages 10–17.

Arthur C. Graesser and Natalie K. Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

Bert F. Green, Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Proceedings of the 1961 Western Joint Computer Conference, (IRE-AIEE-ACM 1961)*, pages 219–224.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 5003–5009.

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 3023–3030.

Eduard Hovy, Ulf Hermjakob, and Deepak Ravichandran. 2002. A question/answer typology with surface text patterns. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT 2002)*, pages 247–251.

Eduard H. Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, and Chin-Yew Lin. 2000. Question answering in Webclopedia. In *Proceedings of The Ninth Text REtrieval Conference (TREC 2000)*.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM 2019)*, pages 105–113.

Ryu Iida, Canasai Kruengkrai, Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Exploiting background knowledge in compact answer generation for why-questions. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019), The Thirty-First Innovative Applications of Artificial Intelligence Conference (IAAI 2019), The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2019)*, pages 142–151.

Ryo Ishida, Kentaro Torisawa, Jong-Hoon Oh, Ryu Iida, Canasai Kruengkrai, and Julien Kloetzer. 2018. Semi-distantly supervised neural model for generating compact answers to open-domain why questions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, pages 5803–5811.

Mandar Joshi, Eunsol Choi, Daniel S. 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 (ACL 2017)*, pages 1601–1611.

Humayun Kayesh, Md. Saiful Islam, and Junhu Wang. 2019. On event causality detection in tweets. *CoRR*, abs/1901.03526.

Humayun Kayesh, Md. Saiful Islam, Junhu Wang, Shikha Anirban, A. S. M. Kayes, and Paul A. Watters. 2020. Answering binary causal questions: A transfer learning based approach. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2020)*, pages 1–9.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. 2022. UnifiedQA-v2: Stronger generalization via broader cross-format training. *CoRR*, abs/2202.12359.

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

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics (EMNLP 2021)*, pages 421–433.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

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

Wendy Grace Lehnert. 1977. *The Process of Question Answering*. Ph.D. thesis, Yale University.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 7871–7880.

Patrick S. H. Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Trans. Assoc. Comput. Linguistics*, 9:1098–1115.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021), Thirty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI 2021), Eleventh Symposium on Educational Advances in Artificial Intelligence, (EAAI 2021)*, pages 13480–13488.

Xiaolu Lu, Soumajit Pramanik, Rishiraj Saha Roy, Abdalghani Abujabal, Yafang Wang, and Gerhard Weikum. 2019. Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*, pages 105–114.

John McClure, Denis J Hilton, Jodie Cowan, Lucyna Ishida, and Marc Wilson. 2001. When people explain difficult actions, is the causal question how or why? *Journal of Language and Social Psychology*, 20(3):339–357.

Arindam Mitra and Chitta Baral. 2016. Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI 2016)*, pages 2779–2785.

Salman Mohammed, Peng Shi, and Jimmy Lin. 2018. Strong baselines for simple question answering over knowledge graphs with and without neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pages 291–296.

Dan I. Moldovan, Sanda M. Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, and Vasile Rus. 2000. The structure and performance of an open-domain question answering system. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 563–570.

Dan I. Moldovan, Marius Pasca, Sanda M. Harabagiu, and Mihai Surdeanu. 2003. Performance issues and error analysis in an open-domain question answering system. *ACM Trans. Inf. Syst.*, 21(2):133–154.

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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*.

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 (ACL 2018)*, pages 784–789.

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 (EMNLP 2016)*, pages 2383–2392.

Apoorv Saxena, Aditay Tripathi, and Partha P. Talukdar. 2020. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 4498–4507.

Roger C. Schank. 1975. The structure of episodes in memory. In *Representation and Understanding*, pages 237–272. Morgan Kaufmann.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1073–1083.

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. 2016. Creating causal embeddings for question answering with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 138–148.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4149–4158.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. 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 2019)*, pages 6075–6084.

Robert S Taylor. 1962. The process of asking questions. *American Documentation*, 13(4):391–396.

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 (Rep4NLP@ACL 2017)*, pages 191–200.

Suzan Verberne, Lou Boves, Peter-Arno Coppen, and Nelleke Oostdijk. 2006a. Discourse-based answering of why-questions. *Trait. Autom. des Langues*, 47(2):21–41.

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

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007)*, page 735–736.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2008. Using syntactic information for improving why-question answering. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 953–960.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2010. What is not in the bag of words for *Why*-QA? *Comput. Linguistics*, 36(2):229–245.

Daniel Vollmers, Rricha Jalota, Diego Moussallem, Hardik Topiwala, Axel-Cyrille Ngonga Ngomo, and Ricardo Usbeck. 2021. Knowledge graph question answering using graph-pattern isomorphism. In *Proceedings of the 17th International Conference on Semantic Systems (SEMANTiCS 2017)*, pages 103–117.

Michael Völske, Pavel Braslavski, Matthias Hagen, Galina Lezina, and Benno Stein. 2015. What users ask a search engine: Analyzing one billion russian question queries. In *24th ACM International Conference on Information and Knowledge Management (CIKM 2015)*, pages 1571–1580.

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

Weiguo Zheng, Jeffrey Xu Yu, Lei Zou, and Hong Cheng. 2018. Question answering over knowledge graphs: Question understanding via template decomposition. *Proc. VLDB Endow.*, 11(11):1373–1386.