A Non-Factoid Question-Answering Taxonomy

Valeriia Bolotova
lurunchik@gmail.com
RMIT University

Vladislav Blinov
vladislav.a.blinov@gmail.com
Ural Federal University

Falk Scholer
falk.scholer@rmit.edu.au
RMIT University

W. Bruce Croft
croft@cs.umass.edu
University of Massachusetts Amherst

Mark Sanderson
mark.sanderson@rmit.edu.au
RMIT University

ABSTRACT

Non-factoid question answering (NFQA) is a challenging and under-researched task that requires constructing long-form answers, such as explanations or opinions, to open-ended non-factoid questions – NFQs. There is still little understanding of the categories of NFQs that people tend to ask, what form of answers they expect to see in return, and what the key research challenges of each category are.

This work presents the first comprehensive taxonomy of NFQ categories and the expected structure of answers. The taxonomy was constructed with a transparent methodology and extensively evaluated via crowdsourcing. The most challenging categories were identified through an editorial user study. We also release a dataset of categorised NFQs and a question category classifier[1].

Finally, we conduct a quantitative analysis of the distribution of question categories using major NFQA datasets, showing that the NFQ categories that are the most challenging for current NFQA systems are poorly represented in these datasets. This imbalance may lead to insufficient system performance for challenging categories. The new taxonomy, along with the category classifier, will aid research in the area, helping to create more balanced benchmarks and to focus models on addressing specific categories.

CCS CONCEPTS

• **Information systems → Query intent**; *Document structure*; **Presentation of retrieval results**; **Question answering**; **Clustering and classification**; **Answer ranking**; • **Computing methodologies → Language resources**; **Supervised learning by classification**; **Cluster analysis**.

KEYWORDS

non-factoid question-answering, question taxonomy, dataset analysis, editorial study

ACM Reference Format:
Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A Non-Factoid Question-Answering Taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531926

1 INTRODUCTION

The task of question answering (QA) is to return an answer to a natural language question. Research into factoid QA has been highly successful, and included the development of various large-scale datasets such as SQuAD [34] and MS MARCO [32], as well as the implementation of transformer-based models such as ALBERT [25] that are able to exceed human performance. However, much less research has been conducted for non-factoid question answering (NFQA), where longer passage-level answers such as opinions or explanations are expected. The performance of state-of-the-art systems on existing datasets such as NFL6, ANTIQUE, NLQuAD and ELI5 [11, 14, 21, 39] falls far behind that of humans [14, 24, 39]. Moreover, QA systems in industry, including answer snippet generation on search engine results pages (SERPs) or conversational agents, are still unlikely to be able to meaningfully answer NFQs such as *"If scientifically possible, should humans become immortal?"*.

Even when systems reach a point where they are able to perform well on existing datasets, there is no guarantee that they will generalize to all categories of NFQs, especially more complex categories such as ones that require a summary of multiple points of view or experiences. In fact, no analysis has been conducted on the distribution of question categories in NFQA datasets, beyond considering starting question words. There is thus a risk that current NFQA systems ignore under-represented categories, focusing on popular and simpler cases.

While we believe that the ultimate goal of NFQA could be an end-to-end system that can deal with all categories of NFQs, it would be beneficial at this stage to study each question category separately, focusing on their unique challenges. Namely, it may be more efficient to use different generative algorithms to construct specific answer structures for each category. Consider two example NFQs: *"How to come up with ideas?"* and *"What is the meaning of nkg?"*. For the first question, the answer should contain a description of the process with concrete steps for different approaches. In contrast, the answer to the second question should list all definitions of the abbreviation with necessary explanations and examples. Given that the expected answer structure is different across question categories, we believe that it is important to understand which

[1] https://github.com/Lurunchik/NF-CATS

SIGIR '22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531926

answer structures are needed and what their unique challenges are. To do so, we must first define possible question categories.

Unfortunately, there is no unified and well-evaluated taxonomy for NFQs, unlike factoid QA where a few taxonomies of question categories and forms of target answers exist [22, 28, 41]. While some related works (described in Section 2) involve taxonomies of NFQ categories, the information on particular details of those taxonomies is rather scattered. In our preliminary user study, described in Section 3.2, we tried to adopt an existing theoretical taxonomy [42] for complex questions, but the agreement on question categories between study coordinators was extremely poor, and did not improve even after a few rounds of discussion. For example, there is only a nuanced difference between the Causal Antecedent and Causal Consequent categories of that taxonomy. Thus, there was a need to gather information on existing NFQ taxonomies from all available sources and to create a taxonomy that is built with a transparent methodology and is thoroughly evaluated.

In this paper, we aim to accelerate the research of non-factoid QA by studying which categories of NFQs exist, what their distribution is in existing datasets, and what potential forms of answers they require. Our contributions can be summarized as follows:

- We propose a new taxonomy of NFQ categories and their respective target answer structures. We revised the initial taxonomy version via a controlled editorial user study. The study also revealed which categories are the most difficult to answer from a human perspective, and how system- and human-generated answers for different categories compare. We extensively evaluated the taxonomy via crowdsourcing studies, including a comparison of how people group questions naturally, when no taxonomy is provided. (Section 3)
- We release[1] a dataset of NFQ categories along with a well-performing model for category classification. (Section 4)
- We provide an analysis of NFQ category distributions in various public QA datasets and evaluate the per-category performance of a state-of-the-art NFQA model. (Section 5)

2 RELATED WORK

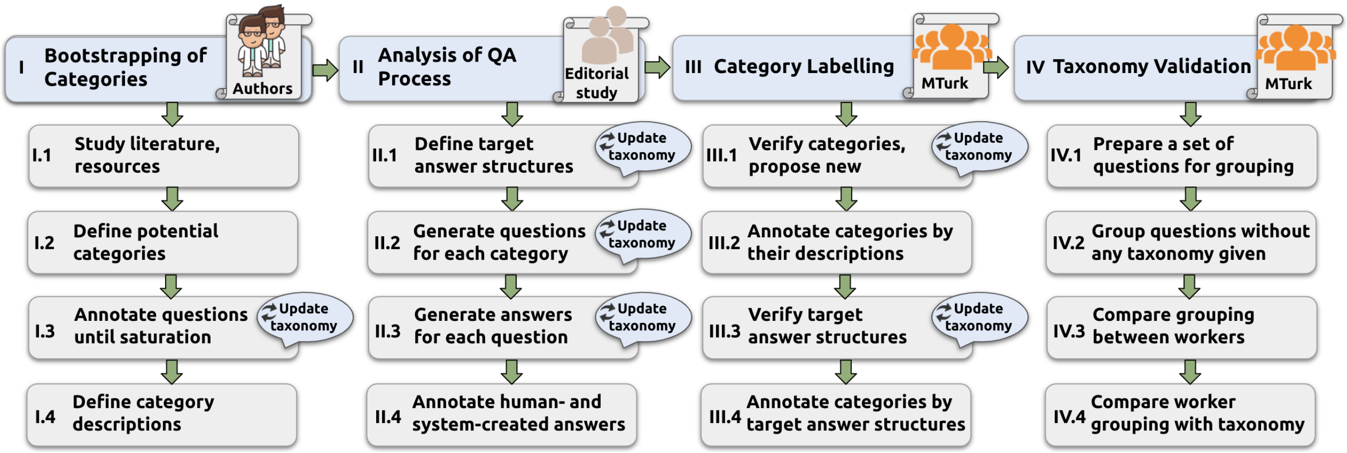
In 1977, Lehnert [27] stated that to answer naturally asked questions, a system needs a theory of how people ask questions and what answer types are expected. She proposed 13 conceptual categories such as Goal Orientation, Instrumental/Procedural, Quantification, Verification, etc. Graesser and Person [17] later extended the taxonomy with five categories. Burger et al. [6] called for new question taxonomies, highlighting limitations of past taxonomies such as a lack of scalability for the larger scope of open-domain QA, and no actual implementations of these taxonomies due to their usage requiring question processing based on a specific knowledge representation. Chaturvedi et al. [8] underline that the creation of a reusable taxonomy for non-factoid questions is a difficult task requiring considerable manual efforts and is expensive.

The Text Retrieval Conference (TREC 23) QA track evaluated systems that answer factual questions. Question category classification was an important component of such systems [44] and factoid question taxonomies emerged [20, 38, 40]. These were all based on the target answer form: person, location, date, etc. However, there was no standard hierarchy of question types. Hovy et al. [22] created a QA topology of 140 answer types by manually analysing 17,000 questions, including categories proposed for NFQs (narrative answer types). The latter were marked as tentative by the authors,

Table 1: The proposed taxonomy of NFQ categories and target answer structures

Category	Description	Expected Answer Structure	Patterns
INSTRUCTION	You want to understand the **procedure/method** of doing/achieving something.	Instructions/guidelines provided in a step-by-step manner.	How to ...? How can I do . . . ? What is the process for . . . ? What is the best way to . . . ?
REASON	You want to find out **reasons** of/for something.	A list of reasons with evidence.	Why does . . . ? What is the reason for . . . ? What causes . . . ? How come ... happened?
EVIDENCE-BASED	You want to learn about the **features/description/definition** of a concept/idea/object/event.	Wikipedia-like passage describing/defining an event/object or its properties based only on facts.	What is . . . ? How does/do . . . work? What are the properties of . . . ? What is the meaning of . . . ? How do you describe . . . ?
COMPARISON	You want to **compare/contrast** two or more things, understand their differences/similarities.	A list of key differences and/or similarities of something compared to another thing.	How is X . . . to/from Y? What are the . . . of X over Y? How does X . . . against Y?
EXPERIENCE	You want to **get advice** or **recommendations** on a particular topic.	Advantages, disadvantages, and main features of an entity (product, event, person, etc) summarised from personal experiences.	Would you recommend . . . ? How do you like . . . ? What do you think about . . . ? Should I . . . ?
DEBATE	You want to **debate on a hypothetical question** (is someone right or wrong, is some event perceived positively or negatively?).	Arguments on a debatable topic consisting of different opinions on something supported or weakened by pros and cons of the topic in the question.	Does . . . exist? Can . . . be successful? Do you think . . . are . . . ? Is . . . really a . . . ?

Figure 1: Taxonomy creation procedure


I Bootstrapping of Categories
Authors
II Analysis of QA Process
Editorial study
III Category Labelling
MTurk
IV Taxonomy Validation
MTurk
I.1 Study literature, resources
I.2 Define potential categories
I.3 Annotate questions until saturation
Update taxonomy
I.4 Define category descriptions
II.1 Define target answer structures
Update taxonomy
II.2 Generate questions for each category
Update taxonomy
II.3 Generate answers for each question
Update taxonomy
II.4 Annotate human- and system-created answers
III.1 Verify categories, propose new
Update taxonomy
III.2 Annotate categories by their descriptions
III.3 Verify target answer structures
Update taxonomy
III.4 Annotate categories by target answer structures
IV.1 Prepare a set of questions for grouping
IV.2 Group questions without any taxonomy given
IV.3 Compare grouping between workers
IV.4 Compare worker grouping with taxonomy

who suggested that more work needs to be done to evaluate and finalize those types. Suzuki et al. [41] proposed another taxonomy for factoid questions which was a hierarchy of 150 categories derived from analysis of about 5,011 questions in Japanese.

Li and Roth [28] proposed a widely adopted two-layer taxonomy consisting of six general question categories and fifty subcategories, and published a dataset of 6,000 labelled questions. Their taxonomy was focused primarily on factoid questions, only partially covering NFQs in a single DESC category (description and abstract concepts) represented by 1,286 questions. In our work, we expand this category to cover all NFQs, provide more detailed descriptions of categories and target answers, and conduct a thorough evaluation. Relying on these categories while creating an initial draft of the new taxonomy, we did not use the data the authors provided in our editorial studies and model training process in order to fairly compare the two taxonomies. In Section 5 we provide an analysis of our question category prediction model on this dataset, and map categories from one taxonomy to the other.

Gupta et al. [18] extended Li and Roth's taxonomy through inclusion of additional sub-categories for DESC: CAUSE & EFFECT, COMPARE AND CONTRAST, and ANALYSIS. Since the NFQA taxonomy was not the focus of their paper, the authors provide only one example question for each category, they omit descriptions of new categories or expected forms of answers, do not give enough information on the methodology used to establish the categories, and do not evaluate the taxonomy to identify overlapping categories and question coverage. We drew some inspiration from their categorization during the bootstrapping stage explained in Section 3.1 and in the process found a few discrepancies such as REASON being a subset of CAUSE & EFFECT, and DESCRIBE and ANALYSIS having a blurry boundary between them.

Motivated by the Broder [4] taxonomy of user intents (later extended by Rose and Levinson [35]), Bu et al. [5] proposed a function-based QA taxonomy obtained through manual analysis of questions asked on Baidu Zhidao. It consists of six categories: fact, list, reason, solution, definition, and navigation. Unlike Broder who surveyed Alta Vista searchers along with an analysis of system logs, Bu et al. did not detail how final categories were chosen or present an evaluation of this taxonomy in terms of consistency and clarity. Mizuno et al. [31] proposed a categorization of NFQs based on types of expected answers, and annotated 2,064 randomly sampled QA pairs from a Japanese CQA platform using that categorization. A detailed description and evaluation of the categories were not provided by the authors. Verberne et al. [43] studied the WHY NFQ category, analyzing syntactic forms of questions and types of answers that fall within that category. They released a dataset of WHY questions.

Leveraging archives of question-answering data from Yahoo!Answers, Chen et al. [9] and Guy et al. [19] categorized possible intents of CQA users. Chen et al. identified objective, subjective, and social intents, while Guy et al. classified them into informational or conversational. The broad nature of these taxonomies and focus on social aspects rather than on the form of questions and answers complicates their direct application in NFQA.

3 NON-FACTOID QA TAXONOMY

The proposed taxonomy of NFQ categories and target answer structures was created through an iterative process shown in Figure 1. The final taxonomy, with examples, is given in Table 1. In this section, we describe each step of taxonomy creation and verification, and explore the most difficult-to-answer categories.

3.1 Bootstrapping of Categories

To create a draft of the NFQ taxonomy, an initial set of categories came from studying the literature on NFQA and researching web-resources dedicated to specific categories of NFQs such as wikihow.com, debate.org, and diffen.com. First, the authors identified a set of disjoint question categories covering most of the questions. Then, a set of questions was assessed by the authors to refine the categories and evaluate the comprehensiveness of the preliminary taxonomy. Questions were randomly sampled from each of the datasets NFL6 [11], MS MARCO [32], ELI5 [14], PhotoshopQuiA [13], SubjQA [3], StackExchange, and Quora Question Pairs and crawled dataset from kialo.com. As we focused on self-contained NFQs (such as web-search queries), we did not include the NLQuAD dataset where most questions depend on external context in order to be understood and disambiguated [39]. In total, the authors assessed 800 questions before reaching saturation, as no new categories or ambiguities emerged further in the assessment.

3.2 Analysis of QA Process

To determine target answer structures and understand the answer generation complexity for each category, we conducted an editorial user study. The study was reviewed and approved by the Human Research Ethics Committee of the RMIT University. All communication with the participants was online, and all participants signed a consent form, declaring that they fully understood the purpose of the study and agreed that their anonymized data can be used. Each participant was compensated with a gift voucher valued at $500 (AUD). All twelve participants were fluent in English: ten native speakers, and two proficient speakers. The mean participant age was 32, ranging from 18 to 57, ten participants were female. Eight participants had at least a bachelor's degree, and four had at least a high school degree. All of them were active computer users. The expertise areas of the participants were diverse, including linguistics, arts, and economics.

During discussions with study participants, we further refined the taxonomy by establishing target answer structures and reformulating category descriptions based on clarification questions asked by the participants. Some categories were renamed to better reflect the target answer structures. The reasons were twofold: first, to align our taxonomy more closely with factoid question taxonomies, where categories are typically named after the type of entity expected in the answer: PERSON, LOCATION, NUMBER, etc; and second, to avoid biasing annotators towards expecting that questions of certain categories should start with specific words (hence renaming HOW-TO to INSTRUCTION and WHY to REASON).

The study consisted of three stages:

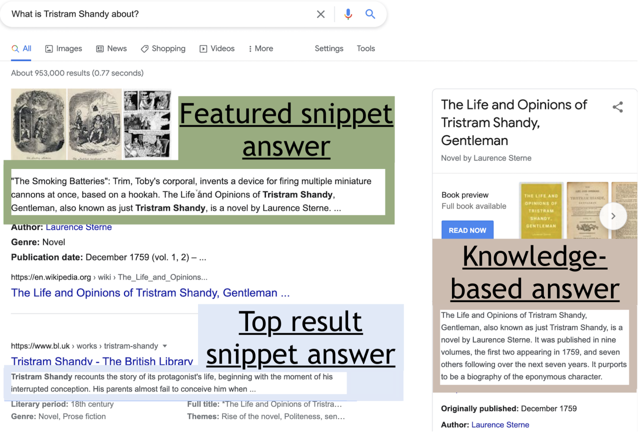
(1) question generation and labeling;
(2) answer generation;
(3) answer labeling.

To review the study procedure and new taxonomy description, the authors performed a test run of the study with three study coordinators during which some explanatory changes were made, and preliminary expected answer structures for each category were created.

Stage 1: Each participant acted as a question asker, and generated a number of questions for each category of the new taxonomy. The categories were explained to the askers in detail, together with a sample set of question patterns for each category. They were then asked to come up with four questions for each NFQ category, as well as four factoid questions to enable comparative analysis (28 questions in total). For the FACTOID category, participants were given the following description: "You expect to see a short sentence containing a word, a short phrase, or an entity name as the answer. (What is the name of . . . ? Where is . . . ? When is . . . ? Who is . . . ? Whose . . . is . . . ?)".

Participants were encouraged to use their recent search history (i.e., queries they previously submitted to a web search engine), questions from CQA websites, etc. For each question, askers were

Figure 2: Three types of Google SERP snippets available for the question "What is Tristram Shandy about?"



About 953,000 results (0.77 seconds)

"The Smoking Batteries": Trim, Toby's corporal, invents a device for firing multiple miniature cannons at once, based on a hookah. The Life and Opinions of **Tristram Shandy**, Gentleman, also known as just **Tristram Shandy**, is a novel by Laurence Sterne. ...

Author: Laurence Sterne
Genre: Novel
Publication date: December 1759 (vol. 1, 2) – ...

https://en.wikipedia.org › wiki › The_Life_and_Opinions...
The Life and Opinions of Tristram Shandy, Gentleman ...

https://www.bl.uk › works › tristram-shandy
Tristram Shandy - The British Library
Tristram Shandy recounts the story of its protagonist's life, beginning with the moment of his interrupted conception. His parents almost fail to conceive him when ...

Literary period: 18th century
Full title: "The Life and Opinions of Tristra...
Genre: Novel, Prose fiction
Themes: Rise of the novel, Politeness, sen...

The Life and Opinions of Tristram Shandy, Gentleman
Novel by Laurence Sterne

The Life and Opinions of Tristram Shandy, Gentleman, also known as just Tristram Shandy, is a novel by Laurence Sterne. It was published in nine volumes, the first two appearing in 1759, and seven others following over the next seven years. It purports to be a biography of the eponymous character.

Originally published: December 1759
Author: Laurence Sterne

also requested to assess the difficulty level of their questions using a three-point scale, to indicate how likely they think it would be for a human to provide a useful answer after consulting information sources, in other words, how difficult it is from the askers' perspective to satisfy their information need. The perceived question difficulty is shown in the first column of Table 2.

Stage 2: Each set of 28 questions generated by an asker was presented to another participant (answerer), who was requested to find answers to each question. There was no direct communication between askers and answerers, participants received their tasks through the study coordinator. The answerers could consult any information source and provide answers of any length as long as they were potentially useful to the asker. The final answers had to be written in the answerer's own words as participants were not allowed to copy-and-paste from another source.

In addition to human answers, "system" answers were retrieved from Google web-search snippets for each of the 336 questions generated by the participants. The questions were submitted as queries, and then the answers were extracted from the snippets in the following priority, whichever was the first available:

(1) FEATURED SNIPPET ANSWER appearing top on Google's SERP;
(2) KNOWLEDGE-BASED ANSWER appearing on the right;
(3) TOP RESULT SNIPPET ANSWER extracted from the most relevant search result and are always present.

Example snippets of each type are shown in Figure 2. In total, 57% system answers came from FEATURED SNIPPET ANSWER, 32% from TOP RESULT SNIPPET ANSWER, and 11% from KNOWLEDGE-BASED ANSWER.

Stage 3: Both human- and system-generated answers were returned to the askers, who then assessed the quality of each answer; the askers were not given information about the answer source. Table 3 provides examples of generated questions and answers for each category along with quality assessments from the askers.

Table 2 shows the results of the evaluation study for human- and system- created answers for each category. The system performance varies greatly between the categories. From an asker's perspective, FACTOID and EVIDENCE-BASED questions were expected to be the easiest to answer after consulting information sources ("Question difficulty" column). System-generated answers (Google SERP snippets) for these two categories were judged more useful than human-generated ones. The DEBATE, EXPERIENCE, and REASON questions were judged the most difficult and least likely to receive a useful answer from an asker's perspective, and the usefulness of system-generated answers for these categories was judged lower than that of human-generated answers. The difference between system- and human-generated answer usefulness was statistically significant (paired t-test, $p < 0.05$) for the DEBATE category.

3.3 Category Labelling

We next evaluated the taxonomy with a series of larger-scale crowdsourcing editorial studies on Amazon Mechanical Turk (MTurk). For all crowdsourcing tasks (HITs), workers were selected following best practices for data collection on MTurk: HIT approval >95%, HITs approved >100 [33]. After several trial tasks for each study, where workers were able to provide feedback regarding the clarity of the task or any concerns about the reward amount, the best-performing workers were selected and assigned a special qualification type for them to continue the actual editorial study, to ensure that it was completed by more reliable assessors. Workers were allowed to participate in only one type of a crowdsourced study, to exclude possible bias from already being familiar with the taxonomy. Each HIT was assessed by 3 workers in all experiments.

Question category labelling: First, to verify how well people understand the question categories based on their descriptions, and whether they have issues with choosing a category, we designed a crowdsourcing study for NFQ category labelling. In each HIT, workers were shown three questions and seven categories: six from the proposed taxonomy plus one additional OTHER/MULTI category for questions that do not fall into any other category, or fall into multiple categories. The categories appeared under each question in a random order, to prevent potential positioning bias. Workers were instructed to read each category description given in the instructions, and choose the most appropriate category for each question. For convenience, a shortened version of the category description appeared when the respective category name was clicked on. For the OTHER/MULTI category, workers were given the option to provide their own category name/description. Until pressing the submit button, workers were free to change the chosen category for each question. During trial annotation runs, all cases where a question received three different labels or suggestions for OTHER/MULTI category were reviewed by the authors, and the necessary changes were made to improve the taxonomy and category descriptions.

Table 2: Perceived difficulty of questions on scale from 0 (very likely to answer) to 2 (not likely to answer); and the usefulness of corresponding human and system answers, rated from 0 (not useful) to 4 (very useful)

Category	Question difficulty	Answer usefulness	
		system	*human*
INSTRUCTION	0.27	2.44	2.40
REASON	0.54	2.15	2.69
EVIDENCE-BASED	0.15	2.71	2.29
COMPARISON	0.29	2.40	2.54
EXPERIENCE	0.50	1.96	2.42
DEBATE*	1.00	1.62	2.29
FACTOID	0.02	3.02	2.69

* *significant difference between system/human answers.*

One out of three questions in each task was a gold question with a known answer, previously assessed by the authors and also given the same label by ≥ 2 workers during trial runs. We had 273 gold questions. A HIT was automatically approved when the gold question was answered correctly, and rejected otherwise. We manually studied all automatic rejections and refunded them if workers had good justifications for their annotation in the rejection form. Each HIT was rewarded with $0.2 upon approval. At the end of this stage, some question category names and descriptions were simplified and clarified based on worker comments.

In total, 1000 questions were assessed by at least three workers in this study, and the final question category was chosen by majority voting. The inter-annotator agreement between assessors was moderate with 0.54 Fleiss' kappa [15], which should also be interpreted in the context of the relatively high number of categories.

To additionally evaluate the quality of assessment and comprehensibility of categories, pairs of duplicate questions from the Quora Question Pairs dataset were mixed into different HITs, based on the assumption that two questions which simply paraphrase each other should belong to the same category. Only 4 pairs out of 154 received inconsistent labels, demonstrating that workers generally agreed on categories for paraphrased questions.

Target answer structure evaluation: To verify how well the expected answer structure describes its corresponding category, we ran another round of crowdsourcing assessments using the same interface, except showing target answer structure descriptions instead of category descriptions. After completing a number of trial runs and selecting assessors, we refined the target answer structures based on workers' comments on task clarity. We used the same approval process as in the previous study, based on gold questions and subsequent manual inspection of rejections. At least three workers assessed 850 questions, with inter-annotator agreement being moderate (0.53 Fleiss' kappa), showing that both category and answer descriptions are equally understandable and suitable for assessment.

In total, for both studies, 12.5% of HITs were automatically rejected, of which 5% had good justifications and were ultimately refunded. Only 7 questions were labelled as the MULTI category, falling under two or more categories; 8 questions questions received 3 different labels 7 of which were assigned MULTI label. The final taxonomy category names, descriptions, and target answer structures are given in Table 1.

3.4 Taxonomy Validation

To study how well our taxonomy corresponds to how people naturally group questions, we asked crowdsource workers to group a set of questions by type or the expected form of an answer. They were not given any guidance (e.g. a taxonomy, or examples) and were

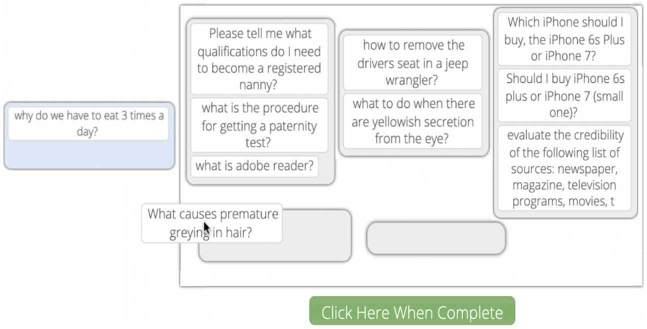
Table 3: Examples of evaluated questions and human/system answers from the editorial user study

Category	Generated question	System answer	Human answer
INSTR.	How can I change the link colour in Elementor? (*Easy to answer: 4.0*)	Select a section and then switch to style. Now you see an option typography. Just click to open that. And now you notice you have an option link color right here. (*Perceived usefulness: 3.0*)	Click on the 3 lines on the top left-hand corner, click "Theme Style" and then click "Typography". Then go to the "Link" heading and from there you can change its colour! (*Perceived usefulness: 3.0*)
REASON	Why are my Peace Lily flowers turning brown? (*Easy to answer: 4.0*)	Peace lily blossoms change color as they age. ... After about 10 days, peace lily blossoms turn green again. Pollination and seed setting also cause the flowers to turn green. The flowers usually stay on the plant for about a month before they begin to die and turn brown, and finally they wither and fade. (*Perceived usefulness: 4.0*)	All flowers die but you are probably may be over or under watering. Most likely overwatering. (*Perceived usefulness: 3.0*)
EVID.	What are the benefits of topically applying Vitamin E? (*Easy to answer: 4.0*)	Vitamin E is a powerful antioxidant that may be effective at reducing UV damage in skin. And vitamin E applied topically may help nourish and protect your skin from damage caused by free radicals. (*Perceived usefulness: 4.0*)	It can repair skin and lighten dark spots but you can also have too much of it. (*Perceived usefulness: 1.0*)
COMP.	How are organic vegetable seeds better than regular seeds? (*Easy to answer: 4.0*)	Organic Seeds Are More Robust Seeds and plants do not know they are being grown organically. ... The genetics of the seed does not change after growing plants organically for a few years. Good quality seeds from organic farms or conventional farms will grow equally well in your soil. (*Perceived usefulness: 4.0*)	There really is no difference between them! It's mostly marketing spin. (*Perceived usefulness: 0.0*)
EXP.	Where is the best Italian restaurant in [LOC]?[2] (*Easy to answer: 2.0*)	Best Italian Restaurant in [LOC] - Menu, Photos, Ratings and Reviews of Restaurants serving Best Italian in [LOC]. Best [LOC] Italian. (*Perceived usefulness: 0.0*)	Da Noi is rated 4/6 on Zomato website. (*Perceived usefulness: 4.0*)
DEBATE	Does god exist? (*Easy to answer: 2.0*)	There remain many mysteries that are beyond science. Does that mean that a God truly exists? A scholar gives reasons for this possibility. (*Perceived usefulness: 0.0*)	Many believe he does and many believe he doesn't. It's up to you to make up your mind about whether you believe God exists. (*Perceived usefulness: 1.0*)

free to define the groups as they saw fit. The procedure for choosing reliable assessors, and clarification of the task, was the same as described in Section 3.3. The questions that participants were asked to group came from the set of questions that were assessed during the previous editorial and crowdsourcing studies.

Here, we aimed to imitate the process of category bootstrapping (Section 3.1) on a smaller scale and compare the results to the new taxonomy. We designed an interface, shown in Figure 3, where questions from the blue area on the left were dragged and dropped into the white box on the right to form a new grey box or expand one of existing grey boxes (question groups). Before clicking the submit button, workers could change their grouping arbitrarily. At the beginning of the task, workers saw ten questions to be arranged and no pre-defined gray boxes. Since it might have felt more natural for participants to attempt to group questions based on topics, rather than on question forms as required, we constructed each set of questions in the task so as to include as many different topics as possible, by clustering the whole set of questions in the dataset prior to the study. We chose HDBSCAN [7] as the clustering algorithm due to its robustness to noise. The questions

Figure 3: Interface for the question clustering study



why do we have to eat 3 times a day?

Please tell me what qualifications do I need to become a registered nanny?

what is the procedure for getting a paternity test?

what is adobe reader?

how to remove the drivers seat in a jeep wrangler?

what to do when there are yellowish secretion from the eye?

Which iPhone should I buy, the iPhone 6s Plus or iPhone 7?

Should I buy iPhone 6s plus or iPhone 7 (small one)?

evaluate the credibility of the following list of sources: newspaper, magazine, television programs, movies, t

What causes premature greying in hair?

were featurised using Universal Sentence Encoder (USE) [26]. The parameters “min_cluster_size” and “min_samples” were set to 2 for HDBSCAN, with the default values used for other parameters. We obtained 1261 clusters in total.

For each HIT we randomly sampled eight questions from different clusters, reducing the likelihood of paraphrased questions or similar topics occurring in a set. The remaining two (out of ten) questions in each HIT were previously annotated paraphrased questions from the Quora Question Pairs dataset. A HIT was automatically approved if the following conditions were met:

(1) the number of submitted question groups was > 1;
(2) the two paraphrased questions from Quora were allocated into the same group.

The first check was enabled by ensuring that each question set contains more than one topic and more than one category. The second check was supported by the idea that two paraphrased questions with the same category label should fall into the same group regardless of the grouping logic. Each of the 36 HITs (360 questions/72 gold items) was completed by three workers; only non-gold questions are used in all subsequent analysis. The workers were rewarded $0.5 for each approved HIT.

Inter-participant cluster similarity: To investigate the level of similarity between the questions groups (further referred to as clusters) created by different participants, we study the agreement between workers, framing the clustering problem as a binary classification task [2]. Each pair of questions receives label 1 if these questions were assigned to the same group, and label 0 otherwise. After this transformation, we can calculate the agreement between participants using Fleiss’ kappa; the workers had almost no agreement with a kappa value of 0.05. Given the sophisticated HIT approval process, we attribute the absence of agreement to the task of unsupervised clustering being much more challenging than labelling in accordance with an existing categorization. Manual verification of 150 random groups confirmed that the workers did create logical groupings of questions.

Clustering and taxonomy similarity: To evaluate the similarity between the natural groupings in this task and taxonomy categories, for each question set in a HIT, clusters created by workers in the current experiment can be compared to the “reference” clusters created by aggregating questions using category labels previously assessed by the question classification crowdsourcing study. This is quantified using the V-measure [36], which compares clusters with reference clustering in terms of homogeneity and completeness. In our case, the V-measure score was 0.6 (with homogeneity of 0.73 and completeness of 0.55). Reference clustering based on the taxonomy on average consisted of 3.6 clusters (i.e. on average each set of 8 questions contained 3.6 non-factoid categories) while workers recorded a mean of 5.5 clusters. The relatively high homogeneity score shows that workers usually grouped questions together similarly to our taxonomy categorization, but their clustering was slightly more fine-grained.

The results of the study show that people with no prior knowledge of our question taxonomy naturally tend to place questions of the same taxonomy category together in one group; however, their groups are typically smaller, and vary substantially from person to person. Manual data inspection showed that assessors in our study seemed to be using some additional individual rules for creating subcategories. A random example of grouping performed by MTurk

Table 4: Random examples of questions grouped by workers

	Clusters	
worker #1	1	what is a cultivator? what is adrenogenital syndrome?˙
	2	how do i burn dvd’s using window’s media player? how can i show messege box in web based c#?
	3	How do I offset irregular periods and get back on a regular menstrual cycle? How expensive is it to call finland?
	4	What happened to the Greek Gods and Goddess?
	5	how do you get illeagls to get out of the u.s.?
worker #2	1	what is a cultivator? What happened to the Greek Gods and Goddess?
	2	how do i burn dvd’s using window’s media player? how can i show messege box in web based c#?
	3	how expensive is it to call finland? how do you get illeagls to get out of the u.s.?
	4	What is adrenogenital syndrome? How do I offset irregular periods and get back on a regular menstrual cycle?
worker #3	1	what is a cultivator?
	2	how do i burn dvd’s using window’s media player? how do you get illeagls to get out of the u.s.?
	3	How do I offset irregular periods and get back on a regular menstrual cycle? How expensive is it to call finland? How can i show messege box in web based c#?
	4	What happened to the Greek Gods and Goddess?
	5	what is adrenogenital syndrome?
new NFQA taxon.	1	what is a cultivator? What happened to the Greek Gods and Goddess? how expensive is it to call finland? what is adrenogenital syndrome?
	2	how do you get illeagls to get out of the u.s.?
	3	How do i burn dvd’s using window’s media player? How can i show messege box in web based c#? How do I offset irregular periods and get back on a regular menstrual cycle?

Table 5: Breakdown of NF-CATS dataset

Category	Authors	MTurk	Auto	Total
INSTRUCTION	132	413		545
REASON	119	166		285
EVIDENCE-BASED	325	863		1188
COMPARISON	21	62		83
EXPERIENCE	70	75		145
DEBATE	93	99	1224	1416
FACTOID	34		3822	3856
NOT-A-QUESTION			4466	4466
Total	794	1678	9512	11984

Table 6: Performance of classifier models on NF-CATS test set

Category	LogReg TF-IDF	BERT Base	RoBERTa Base	RoBERTa Squad2.0	Dataset size		
	F1-score	F1-score	F1-score	F1-score	Test	Train	Val
INSTRUCTION	0.856	0.916	0.917	0.943*	113	346	86
REASON	0.849	0.893	0.852	0.893	59	181	45
EVIDENCE-BASED	0.860	0.906	0.918	0.946*	237	761	190
COMPARISON	0.750	0.741	0.815	0.828	15	54	14
EXPERIENCE	0.612	0.566	0.528	0.653	26	95	24
DEBATE	0.911	0.952	0.954	0.957	283	906	227
FACTOID	0.954	0.981	0.980	0.987*	771	2468	617
NOT-A-QUESTION	0.993	0.998	0.998	0.997	893	2858	715
Macro F1-score / Total	0.848	0.869	0.870	0.901*	2397	7669	1918

* *significantly different from the TF-IDF baseline (Student's t-test,* $p < 0.05$*)*

workers is presented in Table 4, with the reference clustering based on the taxonomy displayed at the bottom.

Throughout this section, we have described the long process of careful taxonomy construction and its detailed verification. First, after bootstrapping, the taxonomy was refined in an editorial user study where target answer structures for each category were created and the most challenging categories were identified. According to the findings, people consider DEBATE, REASON and EXPERIENCE questions to be the hardest to answer. System-generated answers for the same categories have the lowest answer usefulness score, falling far behind human-generated answers. A study using question clustering showed that the proposed categorization reflects how people naturally group questions without any guidance. This, alongside moderate agreement between assessors in question categorization in accordance with the taxonomy, gives us confidence that the taxonomy categories are well-defined.

4 QUESTION CATEGORY PREDICTION

The labelled data from the authors and MTurk workers was composed into a dataset called NF-CATS. In this section, we describe this new dataset and how we leveraged it for the task of question category prediction, conducting experiments with different classification models.

4.1 NF-CATS Dataset

The NF-CATS dataset contains examples of natural questions divided into categories from our taxonomy and two supplementary categories: FACTOID (questions that require a short factual answer) and NOT-A-QUESTION (sentences without question intent). The supplementary categories are introduced in order to facilitate the training of models that can predict question categories based on any given text. Questions from these categories were mostly collected via unsupervised means, the details of which are provided below. A large portion of diverse DEBATE questions was also obtained in an unsupervised fashion from kialo.com (a web-resource that specialises in debates). For each question in our dataset, we provide a column that indicates the source of the assessment: (1) *MTurk*: the category was annotated by three MTurk workers and they reached an agreement; (2) *Authors*: the category was annotated by the authors; (3) *Auto*: the category was assigned in an unsupervised fashion based on the question source (e.g. a web-resource that specializes in a particular category).

The breakdown of the categories and the assessment sources is shown in Table 5, and train/validation/test splits of the dataset are presented in Table 6. The dataset is imbalanced, with some categories rarely appearing in the annotated part and others being substantially augmented with unsupervised data. Suitable unsupervised sources of questions for under-represented categories remain to be found. We do not expect the exploratory analysis in Section 5 to be significantly affected by the class imbalance.

Supplementary categories: As the sources of the FACTOID category, we used three QA datasets that mostly contain factoid questions: TweetQA [45], BoolQ [10], and the development split of SQuAD [34]. To ensure that NFQs were not included, we only extracted questions from SQuAD and TweetQA that satisfy two requirements:

(1) the answer contains < 4 words;
(2) Spacy NER found at least one named entity or number/year in the answer.

Given the simplicity of "Yes/No" questions contained in the BoolQ dataset, we randomly sampled 1,500 questions from it with no further filtering. To populate the NOT-A-QUESTION category, sentences were extracted from document contexts of SQuAD and TweetQA questions which had been selected as FACTOID in the previous step. Each context was split into sentences, and all sentences ending with a question mark were excluded.

4.2 Question Category Classification

To explore the performance of different approaches in the task of question category prediction, we use four models. Per-category and macro F1-scores for the models are shown in Table 6.

First, we use logistic regression over *tf-idf* feature vectors as the baseline classification model. The baseline model was implemented in the Scikit-Learn framework. To automate the selection of high-impact hyper-parameters such as regularization strength and vocabulary size, we use the Optuna [1] hyper-tuning framework with macro F1-score on the validation set as the objective. To find the best hyper-parameters, 1000 trials of hyper-tuning search were executed.

We also leveraged three Transformer models: BERT-base [12], RoBERTA-base [29], and RoBERTa-base fine-tuned on SQuAD2.0. The latter was chosen due to its potentially better domain fit for this task. Each Transformer network was followed by 2 feed-forward layers with Mish activation [30] and a classifier layer on top. Cross-entropy was used as the loss function, and AdamW as the optimizer. As our dataset is imbalanced, we applied batch balancing when training Transformer models, sampling N=8 random examples of each class for each mini-batch. The models were implemented using the AllenNLP [16] framework. The training took up to 10 hours on a single NVidia Tesla P100 16GB GPU. Hyper-parameters were selected manually based on the validation loss, with 10 runs, and shared across all models. Weighted F1-scores on the validation set for the best epochs were equal to 0.954 (5th epoch), 0.957 (4th epoch), and 0.958 (6th epoch) for BERT-base, RoBERTA-base, and RoBERTa-SQuAD2.0, respectively.

As expected, Transformer models provide a substantial gain in performance over the simpler linear baseline model, with RoBERTa-SQuAD performing the best. Even though RoBERTa-SQuAD only moderately outperformed BERT-base and RoBERTa-base in terms of macro F1-scores, we found it to have much better generalisation and higher robustness when manually evaluating predictions on different NFQA datasets, as described in Section 5. For instance, BERT-base had a notable skew towards the INSTRUCTION category in its predictions, resulting in many visible false positives.

5 BENCHMARKING CATEGORIES ON QA DATASETS

We first explore how questions in QA datasets are distributed across the categories of the taxonomy and if there is bias towards certain categories. We then perform a per-category evaluation of a recent NFQA system trained on one of these datasets to understand the performance of a SOTA model separately for each question category and how per-category performance of a NFQA sytem corresponds to per-category answer usefulness of SERP snippets (Table 2).

5.1 Analysis of Category Distribution

The left part of table 7 shows the distributions of categories in four datasets, based on predictions of the top performing RoBERTa-SQuAD model from Section 4.2. We evaluate the dataset for question classification compiled by Li and Roth. The original annotation features 78.2% FACTOID questions and 21.8% NFQs, with 13.8% roughly mapping to the EVIDENCE-BASED category from our taxonomy. Model predictions highlight the same predominance of categories that require factual answers, classifying 89.69% of the dataset into either FACTOID or EVIDENCE-BASED. The manual analysis uncovered a few false positives in rare categories, hinting at those categories having an even smaller representation in reality.

To investigate the distribution of questions submitted to web-search engines, we study the MS MARCO dataset [32], which contains more than 1,000,000 user search queries submitted to Microsoft Bing. According to our analysis, most of the queries are either FACTOID or EVIDENCE-BASED, with the INSTRUCTION category being represented to a smaller degree. This suggests that the performance of models displayed on the MS MARCO leaderboard is mainly a reflection of the ability of systems to answer FACTOID and EVIDENCE-BASED questions. It is arguable whether people rarely ask more sophisticated categories of questions due to a smaller need, or simply because they do not expect current systems to answer them. The existence of dedicated web-resources for DEBATE (debate.org, kialo.com), INSTRUCTION (wikihow.com), and COMPARISON (diffen.com) might point towards the second explanation, necessitating further research of under-represented categories.

Finally, we analyse the open-domain NFQA datasets NFL6 and ELI5, both based on data from CQA platforms. In the NFL6 dataset derived from the Yahoo’s Webscope L6 collection, the largest category is INSTRUCTION, followed by EVIDENCE-BASED and REASON. Together with the extremely small representation of INSTRUCTION questions in MS MARCO, this suggests that people prefer to use CQA platforms over web-search engines for various “how-to” questions. The more narrowly focused ELI5 dataset consists of 270K threads from the “Explain Like I’m Five” Reddit sub-forum. Most of the questions in the dataset are from the REASON and EVIDENCE-BASED categories, representing information requests that require explanations. This indicates that the KILT

Table 7: Analysis of dataset distributions (left) and per-category performance of a state-of-the-art model (right)

	Question category distributions in datasets				**A/B human evaluation on ELI5**				
Categories	Li&Roth *TREC*	MS MARCO *Bing queries*	NFL6 *Yahoo answers*	ELI5 (Test set) *Reddit ELI5*	System Performance	Prefer Gold	Prefer System	Both Good	Both Bad
INSTRUCTION	3.47%	6.00%	40.33%	9.80%	30% (6/20)	50% (10)	15% (3)	15% (3)	20% (4)
REASON	2.18%	2.31%	24.85%	45.80%	40% (8/20)	40% (8)	20% (4)	20% (4)	20% (4)
EVIDENCE-BASED	25.29%	34.52%	23.92%	24.20%	45% (9/20)	30% (6)	15% (3)	30% (6)	25% (5)
COMPARISON	0.65%	0.39%	3.42%	1.80%	40% (8/20)	45% (9)	10% (2)	30% (6)	15% (3)
EXPERIENCE	1.23%	0.21%	2.03%	0.60%	20% (4/20)	60% (12)	10% (2)	10% (2)	20% (4)
DEBATE	0.55%	0.96%	2.70%	8.20%	5% (1/20)	75% (15)	0% (0)	5% (1)	20% (4)
FACTOID	64.40%	55.21%	2.57%	5.50%	—	—	—	—	—
NOT-A-QUESTION	2.23%	0.40%	0.19%	4.10%	—	—	—	—	—
TOTAL	5,894	1,026,758	24,512	87,361	30% (36/120)	50% (60)	12% (14)	18% (22)	20% (24)

ELI5 leaderboard primarily reflects the performance of systems on just these two question categories.

5.2 NFQA Model Performance Across Categories

In Section 3.2, Table 2, we evaluated how well NFQ categories are answered by Google web-search snippets, and identified the most challenging NFQ categories to be DEBATE, EXPERIENCE, and REASON. Here, we focus on evaluating the performance of a state-of-the-art model specifically trained to answer NFQs, to understand the influence of the category imbalance in NFQA datasets and to determine challenging categories for the model. For this, we utilize the ELI5 dataset for abstractive long-form QA. Unfortunately, the unsupervised evaluation methodology for long-form NFQA adopted by Fan et al. [14], namely ROUGE score variants, is not representative of the model performance, to the point where randomly selected answers produce higher scores than ground truth answers [24]. Thus, we leave large-scale unsupervised evaluation for future work, along with the research of more suitable metrics for NFQA. Instead, we carried out A/B human evaluation across different NFQ categories in ELI5, following the human evaluation methodology of Krishna et al. [24].

Questions and gold answers for assessment were sourced from the corrected evaluation split of ELI5 provided by Krishna et al. Model answers were generated using the best system of Krishna et al. (with p = 0.9) consisting of a “contrastive REALM” dense retriever and a generator based on the Routing Transformer, the current state-of-the-art model for representing long-range dependencies in sequences via sparse attention and mini-batch *k*-means clustering [37]. Volunteers were asked to select the “better” answer for one question at a time, choosing between gold and system-generated answers presented in random order and without labelling the source. Unlike the original A/B testing setup used by Krishna et al., in “Tie" situations when both answers were equal in their quality (either good or bad), volunteers were instructed to select “Both Good" or “Both Bad" options, respectively. This change allowed us to evaluate the overall percentage of good answers given by the system. In total, we had 5 volunteers and 120 questions, with 20 questions per each NFQ category. All volunteers were English-speaking and had at least a Master's degree. Question categories were assigned through MTurk evaluation in the same manner as described in Section 3.3. The results are presented in the right part of Table 7.

On average, the system answers were preferred only in 12% of cases, which is slightly less than 14% reported for the system by Krishna et al. We attribute this to the difference in distributions of categories between our evaluation and the original evaluation. Krishna et al. randomly sampled questions, and the majority belonged to REASON and EVIDENCE-BASED categories, while we sampled questions uniformly across categories.

The "System Performance" column gives the overall system performance, which measures the percentage of system answers that were either preferred over gold answers or considered equally good. Similarly to the performance of the production-grade system tested in our editorial study (3rd column in Table 2), the most challenging category for the system trained on ELI5 is DEBATE, where the system gave only one good answer out of 20, followed by EXPERIENCE with four good answers. These two categories are poorly represented in the training data. On the other hand, the performance of the ELI5-trained system for the REASON category is relatively good, which could be explained by a very high representation of this question category. This supports our hypothesis that system performance may be affected by unbalanced question categories in training data, especially for more challenging categories.

6 CONCLUSION

Understanding non-factoid question categories asked by users is essential both for the development of successful NFQA systems and for creating reliable NFQA benchmarks. In this work, we present the first streamlined taxonomy of NFQs built with a transparent methodology and evaluated through editorial and crowdsourcing studies. The labelled data was compiled into a new dataset of NFQ categories. To enable researchers to apply these categories to other datasets, we provide a classifier for this purpose. Subsequent analysis of question categories against four existing QA datasets, commonly used in NFQA, demonstrates that these sets have a skewed representation across the taxonomy. The findings indicate a clear need for creating new datasets that cover this more expansive range of categories and new NFQA models capable of dealing with more challenging and previously under-represented categories such as DEBATE and EXPERIENCE.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 2623–2631. https://doi.org/10.1145/3292500.3330701

[2] Javier Artiles, Enrique Amigó, and Julio Gonzalo. 2009. The role of named entities in Web People Search. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, 534–542. https://aclanthology.org/D09-1056/

[3] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. [n.d.]. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020-11). Association for Computational Linguistics, 5480–5494. https://doi.org/10.18653/v1/2020.emnlp-main.442

[4] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (Sept. 2002), 3–10. https://doi.org/10.1145/792550.792552

[5] Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. 2010. Function-Based Question Classification for General QA. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, 1119–1128. https://www.aclweb.org/anthology/D10-1109

[6] John Burger, Claire Cardie, Vinay Chaudhri, Robert Gaizauskas, Sanda Harabagiu, David Israel, Christian Jacquemin, Chin-Yew Lin, Steve Maiorano, George Miller, Dan Moldovan, Bill Ogden, John Prager, Ellen Riloff, Amit Singhal, Rohini Shrihari, Tomek Strazalkowski, Ellen Voorhees, and Ralph Weishedel. 2003. Issues, Tasks and Program Structures to Roadmap Research in Question Answering (QA). In *Document Understanding Conference*. NIST, NIST. https://www.microsoft.com/en-us/research/publication/issues-tasks-and-program-structures-to-roadmap-research-in-question-answering/

[7] Ricardo J. G. B. Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining, 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II (Lecture Notes in Computer Science)*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.), Vol. 7819. Springer, 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

[8] Snigdha Chaturvedi, Vittorio Castelli, Radu Florian, Ramesh M. Nallapati, and Hema Raghavan. 2014. Joint Question Clustering and Relevance Prediction for Open Domain Non-Factoid Question Answering. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. Association for Computing Machinery, New York, NY, USA, 503–514. https://doi.org/10.1145/2566486.2567999

[9] Long Chen, Dell Zhang, and Levene Mark. 2012. Understanding User Intent in Community Question Answering. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12 Companion)*. Association for Computing Machinery, New York, NY, USA, 823–828. https://doi.org/10.1145/2187980.2188206

[10] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 2924–2936. https://doi.org/10.18653/v1/n19-1300

[11] Daniel Cohen and W. Bruce Croft. 2016. End to End Long Short Term Memory Networks for Non-Factoid Question Answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval (ICTIR '16)*. Association for Computing Machinery, New York, NY, USA, 143–146. https://doi.org/10.1145/2970398.2970438

[12] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[13] Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu, and Seokhwan Kim. 2018. PhotoshopQuiA: A Corpus of Non-Factoid Questions and Answers for Why-Question Answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1438

[14] 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*. Association for Computational Linguistics, Florence, Italy, 3558–3567. https://doi.org/10.18653/v1/P19-1346

[15] J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382.

[16] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A Deep Semantic Natural Language Processing Platform. *CoRR* abs/1803.07640 (2018). arXiv:1803.07640 http://arxiv.org/abs/1803.07640

[17] Arthur C. Graesser and Natalie K. Person. 1994. Question Asking During Tutoring. *American Educational Research Journal* 31, 1 (1994), 104–137. https://doi.org/10.3102/00028312031001104 arXiv:https://doi.org/10.3102/00028312031001104

[18] Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. 2018. Can Taxonomy Help? Improving Semantic Question Matching using Question Taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 499–513. https://aclanthology.org/C18-1042

[19] Ido Guy, Victor Makarenkov, Niva Hazon, Lior Rokach, and Bracha Shapira. 2018. Identifying Informational vs. Conversational Questions on Community Question Answering Archives. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 216–224. https://doi.org/10.1145/3159652.3159733

[20] Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. 2000. FALCON: Boosting Knowledge for Answer Engines. In *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000 (NIST Special Publication)*, Ellen M. Voorhees and Donna K. Harman (Eds.), Vol. 500-249. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec9/papers/smu.pdf

[21] Helia Hashemi, Hamed Zamani, and W. Bruce Croft. 2019. Performance Prediction for Non-Factoid Question Answering. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR '19)*. Association for Computing Machinery, New York, NY, USA, 55–58. https://doi.org/10.1145/3341981.3344249

[22] Eduard Hovy, Ul Hermjakob, and Deep Ravichandran. 2002. A question/answer typology with surface text patterns. (01 2002). https://doi.org/10.3115/1289189.1289206

[23] David A. Hull. 1999. Xerox TREC-8 Question Answering Track Report. In *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999 (NIST Special Publication)*, Ellen M. Voorhees and Donna K. Harman (Eds.), Vol. 500-246. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec8/papers/xerox-QA.pdf

[24] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393

[25] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=H1eA7AEtvS

[26] Guang-He Lee and Yun-Nung Chen. 2017. MUSE: Modularizing Unsupervised Sense Embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 327–337. https://doi.org/10.18653/v1/D17-1034

[27] Wendy G. Lehnert. 1977. A Conceptual Theory of Question Answering. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, USA, August 22-25, 1977*, Raj Reddy (Ed.). William Kaufmann, 158–164.

[28] Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002*. https://aclanthology.org/C02-1150/

[29] 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* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[30] Diganta Misra. 2020. Mish: A Self Regularized Non-Monotonic Activation Function. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press. https://www.bmvc2020-conference.com/assets/papers/0928.pdf

[31] Junta Mizuno, Tomoyosi Akiba, Atsushi Fujii, and Katunobu Itou. 2007. Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Realworld Questions. In *Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access, NTCIR-6, National Center of Sciences, Tokyo, Japan, May 15-18, 2007*, Noriko Kando (Ed.). National Institute of

Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings6/NTCIR/71.pdf

[32] 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), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.), Vol. 1773. CEUR-WS.org. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[33] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. 46, 4 (2014), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

[34] 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)*. Association for Computational Linguistics, Melbourne, Australia, 784–789. https://doi.org/10.18653/v1/P18-2124

[35] Daniel E. Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. In *Proceedings of the 13th International Conference on World Wide Web (WWW '04)*. Association for Computing Machinery, New York, NY, USA, 13–19. https://doi.org/10.1145/988672.988675

[36] Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 410–420. https://www.aclweb.org/anthology/D07-1043

[37] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient Content-Based Sparse Attention with Routing Transformers. *Transactions of the Association for Computational Linguistics* 9 (2021), 53–68. https://doi.org/10.1162/tacl_a_00353

[38] Amit Singhal, Steve Abney, Michiel Bacchiani, Michael Collins, Donald Hindle, and Fernando Pereira. 1999. ATT at TREC-8.

[39] Amir Soleimani, Christof Monz, and Marcel Worring. 2021. NLQuAD: A Non-Factoid Long Question Answering Data Set. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1245–1255. https://doi.org/10.18653/v1/2021.eacl-main.106

[40] Rohini Srihari and Wei Li. 2000. A Question Answering System Supported by Information Extraction. In *Sixth Applied Natural Language Processing Conference*. Association for Computational Linguistics, Seattle, Washington, USA, 166–172. https://doi.org/10.3115/974147.974170

[41] Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda. 2003. Question Classification using HDAG Kernel. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*. Association for Computational Linguistics, Sapporo, Japan, 61–68. https://doi.org/10.3115/1119312.1119320

[42] Andrew Tawfik, Arthur Graesser, Jessica Gatewood, and Jaclyn Gishbaugher. 2020. Role of questions in inquiry-based instruction: towards a design taxonomy for question-asking and implications for design. *Educational Technology Research and Development* 68 (01 2020), 1–25. https://doi.org/10.1007/s11423-020-09738-9

[43] Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2006. Data for question answering: The case of why. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/pdf/525_pdf.pdf

[44] Ellen M. Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. In *Proceedings of The Tenth Text REtrieval Conference, TREC 2001, Gaithersburg, Maryland, USA, November 13-16, 2001 (NIST Special Publication)*, Ellen M. Voorhees and Donna K. Harman (Eds.), Vol. 500-250. National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec10/papers/qa10.pdf

[45] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. TWEETQA: A Social Media Focused Question Answering Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5020–5031. https://doi.org/10.18653/v1/P19-1496