An Intent Taxonomy for Questions Asked in Web Search

B. Barla Cambazoglu
RMIT University
Melbourne, VIC, Australia
barla.cambazoglu@rmit.edu.au

Leila Tavakoli
RMIT University
Melbourne, VIC, Australia
leila.tavakoli@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, VIC, Australia
falk.scholer@rmit.edu.au

Mark Sanderson
RMIT University
Melbourne, VIC, Australia
mark.sanderson@rmit.edu.au

Bruce Croft
University of Massachusetts
Amherst, MA, USA
croft@cs.umass.edu

ABSTRACT

This work presents a new, multi-faceted taxonomy to classify questions asked in web search engines based on the question intent, types of entities mentioned, types of question words, and granularity of the expected answer. Built based on the inspection of 1,000 real-life questions issued to a web search engine, the taxonomy reflects the recent search behavior of users and enables deep understanding of user intents, goals, and expected answers. This taxonomy is more fine-grained than previous query taxonomies, and is designed with the ultimate goal of reducing the inherent ambiguity in determining the intent of questions. In addition, we describe the formal procedure for conducting an editorial study of the taxonomy including its evaluation. The adopted procedure aims to increase assessor agreement without incurring too much overhead. Our results demonstrate that, despite being more fine-grained, the proposed intent categories result in higher agreement between assessors compared to an existing, commonly used taxonomy.

KEYWORDS

question intent taxonomy, editorial study, web search engines

ACM Reference Format:
B. Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406522.3446027

1 INTRODUCTION

Detecting the intent of search queries is an important task in information retrieval (IR). Typically, query intent detection models are trained in a supervised manner using ground-truth data composed of query instances whose intent is manually assessed. The trained models are then used to assign an intent category to unseen queries, which may be used as a feature in high-level IR tasks, such as learning-to-rank, query rewriting, or vertical selection.

A fundamental step in building ground-truth data for intent detection is the construction of a taxonomy of possible query intents, i.e., a predetermined set of intent categories which can be assigned to any given query. Creating a meaningful taxonomy requires tackling three challenges. First, most queries should fall into at least one category, i.e., the taxonomy should provide high coverage. Second, queries should not be assigned to multiple categories, i.e., the taxonomy should have little ambiguity. Third, the taxonomy should not lead to a few categories with many queries and many categories with few queries, i.e., a heavily skewed distribution is not desirable, as this may make it more difficult to train intent detection models.

As our first contribution we introduce a novel, multi-faceted taxonomy of questions asked in web search engines. The proposed taxonomy has the following five facets: question intent, entity types mentioned in the question, question word type, entity type of the answer, and the granularity of the answer. Although the question intent is the most important facet in the proposed taxonomy and forms the main focus here, we believe that the remaining facets are also useful to obtain a good characterization of questions.

The need for a new taxonomy is motivated by three observations.

- Existing intent taxonomies for web search queries do not adequately address the coverage issue outlined before [4, 31]. These taxonomies have become less applicable as search engines evolve and provide new functionality, leading to new types of search tasks which were not captured. Our taxonomy provides high coverage as it was devised based on the inspection of a recent sample of search engine logs.
- Most existing taxonomies are specific to simple keyword queries, rather than more complex questions increasingly observed in modern day web search engines [28]. Our taxonomy specifically focuses on questions. In the literature, there are a few question taxonomies [13, 19], but those that exist were developed on small or unrepresentative samples. Our taxonomy has been developed by inspecting real-life questions submitted to a commercial web search engine.
- Query intent taxonomies [31] are likely to lead to skewed intent distributions when applied to questions, as we show in our results. Despite being relatively more fine-grained, our taxonomy results in a less skewed distribution.

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446027

A taxonomy was created through an editorial study that follows a step-by-step procedure, aiming to increase the agreement between assessors, through the iterative refinement of assessments while not incurring excessive overhead. A detailed description of the adopted procedure is provided, and the experience gathered during its execution is shared. We also detail how a recruited group of assessors labeled questions using this procedure.

The key contributions of this paper are as follows:

- We propose a multi-faceted question taxonomy, reflecting recent trends in questions asked in web search engines.
- We describe the formal procedure used to create our taxonomy. We share our experience regarding caveats and pitfalls encountered during the execution of the procedure.
- Based on the created taxonomy, we provide a characterization of questions asked in modern day web search engines.

The main findings of our work are:

- Despite being more fine-grained, the proposed taxonomy leads to higher editorial agreement compared to the popular query intent taxonomy of Rose and Levinson [31], which we use as a baseline. It also results in a meaningful distribution of intent categories, whereas Rose and Levinson's taxonomy places the majority of questions in a single intent category.
- The multi-phase procedure followed in the editorial study leads to higher assessor agreement with relatively little overhead, as the number of phases increases.
- We identify the emergence of certain question intents which did not receive much research attention previously (e.g., questions with calculation/conversion intent). Moreover, we find that retrieving entire web documents is largely redundant, as the information need of almost all questions can be satisfied with a single passage or a shorter piece of text.

The rest of the paper is organized as follows. The formal procedure adopted for the editorial study is described in Section 2. Sections 3 and 4 contain the setting and details of the conducted editorial study, respectively. The resulting multi-faceted question taxonomy is presented in Section 5. The results of the study are analyzed in Section 6. A brief survey of the related work is given in Section 7. The paper is concluded in Section 8.

2 PROCEDURE

The procedure we adopted in our editorial study involves nine consecutive phases, illustrated in Figure 1. The participants are a coordinator (labeled C) and multiple assessors (labeled A).

Phase 1 (Bootstrapping). The purpose of this phase is to bootstrap the labeling procedure by providing a common ground to all assessors. The coordinator first samples a small number of questions to be labeled. They iterate over the sampled questions and try to identify a set of suitable facets for the taxonomy as well as some suitable categories for each identified facet. They then provide a brief description of every facet and category along with a small number of representative questions. A draft taxonomy and draft guidelines are prepared and passed to the assessors.

Phase 2 (Labeling). Using the draft documents, the assessors individually perform a small-scale labeling study on the sampled questions. This phase is not expected to be too time-consuming.

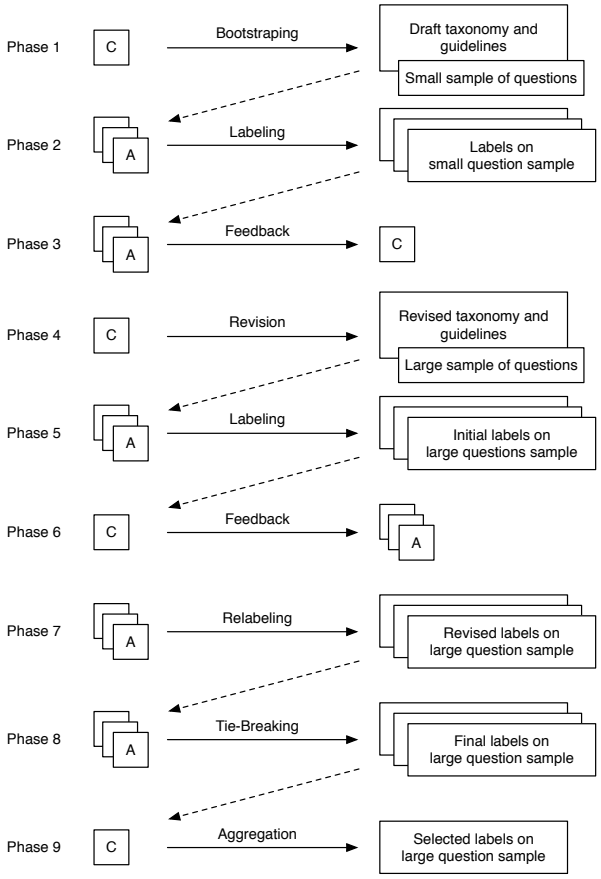

Figure 1: Proposed procedure (letters A and C represent an assessor and the coordinator, respectively).

Phase 3 (Feedback). The coordinator meets with the assessors and receives feedback on challenges encountered. This may include feedback about issues with the current taxonomy, unclear or ambiguous statements in the guidelines, questions that cannot be labeled, and the expected time cost of the study.

Phase 4 (Revision). Based on the assessors' feedback, the coordinator determines if a revised taxonomy is required. If it is, they modify the original guidelines. They also sample a larger set of questions and provide them to the assessors for labeling.

Phase 5 (Labeling). Each assessor independently labels the provided questions based on the revised taxonomy and guidelines. This phase is where the bulk of the work happens.

Phase 6 (Feedback). The coordinator gathers the labels from all assessors and performs an agreement analysis. Here, the goal is to identify the cases where a particular assessor diverged significantly from the rest of the assessors in their assessment. The analysis can be automated by scripts that extract recurring patterns in labels, revealing misinterpretation of guidelines, or other potentially erroneous labeling. Finally, the coordinator provides each assessor with individual feedback, giving an abstract view of cases where they are frequently in disagreement with the remaining assessors and the potential reasons which may have led to disagreement.

Phase 7 (Relabeling). The assessors reconsider their labels in light of the feedback provided by the coordinator in the previous phase.

Although the assessors are expected to review their labels at this stage, they are not obliged to change them if they continue to think that their interpretation is correct.

Phase 8 (Tie-Breaking). Provisional labels are assigned by majority voting. All assessors get together and try to reach agreement on the labels of questions where majority voting does not lead to an outcome. The goal is to resolve all ties and assign a single label to every facet of each question. However, some ties may still remain unresolved as the assessors are not obliged to change their labels.

Phase 9 (Aggregation). The coordinator gathers the final labels of the assessors and aggregates them by assigning a single label to every facet of each question if majority voting leads to a unique label. Otherwise, no label is assigned.

We note that this procedure is applicable to editorial studies where the number of editors is typically small (e.g., less than 10) and each assessor labels every item in the data. Therefore, as long as the labeled data is large enough, the bulk of the work should be in the labeling process, and the phases where coordination takes place are expected to incur a relatively small overhead without becoming a bottleneck. This is also what we have observed in our study, details of which are provided in Section 4.

3 SETTING

Dataset. We base our editorial study on the MS MARCO dataset, which was first made public by Microsoft in 2016 [1] and updated in the following years.[1] The dataset contains around one million anonymized English questions manually selected by human assessors from queries issued to the Bing search engine. The great majority of selected questions are well formed. This dataset has been used in many natural language processing and IR tasks, including passage ranking, question answering, and answer generation.

Assessors. There were five assessors aged from 21 to 45 (one male and four female). The assessors were of diverse ethnicity with different educational backgrounds (one Bachelors, three Masters, and one PhD degree). All assessors were multi-lingual and proficient with the English language although none of them was a native speaker. The coordinator, who also acted as an assessor, had considerable industry experience with conducting similar labeling studies. All assessors completed two ethics courses (Human Research Ethics and Research Integrity) before starting the study as requested by their institution. Throughout the paper, we denote the assessors by arbitrary names (`Ann`, `Beth`, `Cali`, `Dee`, and `Elle`).

Platform and Agreement Measure. Labeling was performed using spreadsheet software. Editorial agreement is computed by means of Krippendorff's α, a chance-corrected agreement measure [15], using the dkpro-statistics library [7].

4 EDITORIAL STUDY

4.1 Phase 1 (Bootstrapping)

The coordinator (`Ann`) came up with a draft taxonomy containing five different facets: question intent, question ambiguity, question topic, expected answer's entity type, and question word type. There were 13 question intent categories. Question ambiguity was a binary facet, indicating whether question was ambiguous or not. The question topics were based on the Open Directory Project's categories[2] while the entity types were taken from the Stanford Core NLP library [24]. Question word types were compiled after inspection of some linguistic resources. The draft guidelines described the facets and provided a list of categories for each facet. In addition, the assessors were asked to label the intent of questions according to Rose and Levinson's taxonomy [31], which would act as a baseline.[3] Finally, the coordinator provided a sample of 50 questions to the assessors, together with the draft taxonomy and guidelines.

4.2 Phase 2 (Labeling)

Assessors individually conducted a small test study on 50 questions, following the provided guidelines. The assessors reported the total time they spent for labeling to be somewhere between 2–4 hours.

4.3 Phase 3 (Feedback)

The coordinator organized a meeting, where the assessors highlighted a number of issues about the draft taxonomy and guidelines:

- Certain questions could not be assigned to any intent category in the draft taxonomy. Thus, the annotators suggested some additional intent categories.
- The annotators also noticed that for some questions, multiple intent categories were applicable at the same time, i.e. some intent categories had overlap, leading to ambiguity.
- Some questions were completely unclear to the assessors, and they had difficulty with labeling those questions. The assessors suggested that it may be better to label the clarity of a question's intent instead of its ambiguity.
- After the discussion, the assessors agreed that the ambiguity of the words in a question do not always lead to multiple intents. For example, in "what is nkg", although the word "nkg" is ambiguous, the intent of the question is simply to obtain the definition of "nkg", which may not be ambiguous to the person who asked the question.
- It was decided that, even if a clearly dominant intent is available, a question should still be labeled as having multiple intents if other intents are also likely.
- Labeling the topic of a question was found to be time-consuming. The topic taxonomy at hand was not suitable, and many questions could not be assigned a topic.
- The entity types used to label the answers were somewhat limited as many named entities could not be covered.
- The assessors raised several concerns about the taxonomy of Rose and Levinson [31]. In particular, the `Obtain` intent category was not very clear to the assessors. They believed that the `Locate` category should have been a subcategory of `Resource` instead of `Informational`, and the `List` category was confusing to some assessors. There were also concerns regarding the ambiguity of some examples in the paper.
- One assessor suggested using an entity tagger to facilitate labeling of the `QuestionEntityType` facet.

[1] The MS MARCO dataset is publicly available at https://microsoft.github.io/msmarco/.

[2] Open Directory Project, https://dmoz-odp.org.

[3] The intent categories available in Rose and Levinson's taxonomy are `Navigational`, `Directed-Closed`, `Directed-Open`, `Undirected`, `Advice`, `Locate`, `List`, `Download`, `Entertainment`, `Interact`, and `Obtain`. The reader may consult the examples in Rose and Levinson's paper [31] to have a better understanding of these categories.

4.4 Phase 4 (Revision)

Based on the feedback received in the previous phase, the coordinator made several substantial changes to the draft taxonomy:

- The question ambiguity facet was removed. Instead, the clarity of questions was assessed using the following labels:
  - `UnclearIntent`: If the intent of a question was not clear to the assessor at all, the question was labeled as `UnclearIntent`. The assessors did not choose this option unless the question was completely unclear to them. Examples of such questions are "asdas dasdasfasfas", "what is normal psi for male pos" (it is not clear what "psi" and "pos" mean even after consulting a search engine), and "25Kg BAG how much area will cover" (the answer depends on the bag's type, which is not clearly stated).
  - `SingleIntent`: This label was assigned when the question had a single, clear intent. Examples are "who sang she wore blue velvet", "how long is canned food safe, "what is parkland near in florida", and "what is nkg".
  - `MultipleIntents`: A question may have multiple possible intents if the user's information need may be interpreted in different ways. This usually happens when the question is too short or under-specified. For example, "tree top resorts gatlinburg" may be trying to locate the resort on a map, book a room there, read some reviews, or simply navigate to its website.
- The question topic facet was removed from the study as a suitable set of topic categories could not be found and creating one from scratch appeared difficult due to the extremely large number of question topics possible in web search.
- `QuestionEntityType` and `AnswerGranularity` were added as two new facets of the taxonomy.
- New intent categories were added (`Advice`, `Opinion`, and `Verification`), bringing the number of categories to 16.
- To solve the overlap issue raised by the assessors, some exclusion rules were added in certain intent categories.
- The list of available entity types was expanded by introducing seven new entity types (`Audio/visual`, `Event`, `Illness`, `Product`, `Fraction`, `Range`, and `Rate`).

The draft guidelines were modified as follows:

- When labeling a question, the assessors were asked to imagine being the user who submitted the question and think about what the user's information need might be. As part of the process, they were allowed to submit the question to a search engine and check the retrieved results to get a better understanding of the user's information need.
- The facets of a question were labeled only if the assessor labeled the `IntentClarity` of the question as `SingleIntent` since it made no sense to perform further labeling for a question which has no clear intent or has multiple intents. Also, if `QuestionIntent` was labeled as `Resource` or `Weather`, the remaining facets were skipped by the assessors since these questions may not seek textual answers.
- Three auxiliary labels were made available for every facet:
  - `None`: This is the default value for all questions at the beginning of the study. It indicates that no choice has been made yet by the assessor for the current question.
  - `?`: Assessors can assign this value to a question if they think none of the available options is suitable for the question.
  - `N/A`: The `N/A` (not applicable) label is assigned when no labeling is required (e.g., `QuestionIntent` is labeled as `N/A` if `IntentClarity` was labeled as `UnclearIntent`).
- The assessors were requested to label questions with a spelling mistake after correcting the mistake if possible.
- The meaning of the `Set` entity type was further clarified.
- Automating the labeling of `QuestionEntityType` facet using an entity tagger was not preferred as a suitable tagger that supports all of our entity types could not be found. However, the assessors were left free to use an entity tagger of their choice to check for presence of entities in questions. They were asked to iterate over such questions and verify the correctness of the tagger's decisions.

The coordinator provided to the assessors 1,000 questions sampled uniformly at random from the MS MARCO dataset. No further cleansing or filtering was performed on questions as the original dataset had already been curated by human assessors. However, one potentially sensitive question was detected and removed from the examples in the guidelines and the paper. Before the next phase began, the coordinator communicated to the assessors that the two assessors with the highest agreement would be given a prize.

4.5 Phase 5 (Labeling)

Each assessor labeled the questions independently, based on the final taxonomy and guidelines they received from the coordinator. All annotators said they labeled the questions by iterating on the facets first, instead of questions. That is, they labeled a particular facet for all questions before moving to a new facet. This approach was found to reduce the overhead of context switching compared to labeling all facets of a question consecutively. `Beth` and `Dee` said that they further sped up the labeling process by grouping questions according to their labels in previously assessed facets (e.g., all questions with `Entity` intent are assessed consecutively), and labeling all questions within a group before moving to the next group. `Elle` used a tool to translate the questions to their native language. The labeling times reported by the annotators ranged between 50–70 hours.

4.6 Phase 6 (Feedback)

The coordinator received the labels provided by all annotators and measured the editorial agreement of each annotator with respect to the remaining annotators in order to reveal potential inconsistencies in the labeling behaviour of assessors. To this end, the coordinator computed the pairwise agreement between each annotator pair using Krippendorff's α measure.

The pairwise average agreement values are shown in the upper section of Table 1. We observe several cases where certain annotators had significantly low agreement with the other annotators (such cases are illustrated in bold): First, `Dee` seems to have somewhat inconsistent labeling of question intent according to Rose and Levinson's taxonomy. Second, `Beth` and `Cali` seem to have difficulty in labeling the entity type of questions as evidenced by almost random agreement values. Third, `Cali` also appears to be inconsistent when labeling the entity type of answers.

Table 1: Average pairwise agreement values associated with each assessor

Phase	Assessors	Intent Clarity	Rose and Levinson	Facets: Question Intent	Question Entity Type	Answer Entity Type	Question Word Type	Answer Granularity
Phase 5 (Labeling)	Ann	0.293	0.293	0.695	0.244	0.658	0.922	0.400
	Beth	0.371	0.243	0.681	**0.024**	0.494	0.904	0.480
	Cali	0.165	0.188	0.591	**-0.026**	**0.198**	0.799	0.361
	Dee	0.276	**-0.010**	0.655	0.230	0.648	0.907	0.414
	Elle	0.383	0.266	0.711	0.282	0.655	0.916	0.521
Phase 7 (Relabeling)	Ann	0.317	0.392	0.723	0.408	0.813	0.961	0.427
	Beth	0.379	0.350	0.710	0.244	0.734	0.944	0.520
	Cali	0.192	0.275	0.667	0.190	0.715	0.936	0.417
	Dee	0.287	0.350	0.675	0.305	0.794	0.941	0.431
	Elle	0.404	0.376	0.728	0.401	0.810	0.951	0.547
Phase 8 (Tie-Breaking)	Ann	0.467	0.455	0.747	0.473	0.820	0.961	0.451
	Beth	0.384	0.359	0.736	0.298	0.738	0.947	0.537
	Cali	0.297	0.329	0.685	0.249	0.719	0.937	0.443
	Dee	0.350	0.368	0.682	0.336	0.796	0.943	0.424
	Elle	0.523	0.442	0.751	0.477	0.814	0.953	0.567

To obtain a list of potential inconsistencies for a given facet, the coordinator selected all questions where four assessors assigned the same label while one assessor had a different assessment. The coordinator then returned every such question to the respective assessor who provided an inconsistent label, for further consideration. The label agreed by the four assessors was not made available to the inconsistent assessor since this could bias their decisions.

4.7 Phase 7 (Relabeling)

After getting feedback about questions where they had seemingly inconsistent labeling, each assessor relabeled their own set of potentially problematic cases. The percentage of labels which were reconsidered by `Ann`, `Beth`, `Cali`, `Dee`, and `Elle` were %0.9, %7.2, %8.2, %6.5, and %0.7, respectively. Thus, the assessors' workload in this phase was relatively low compared to the workload in Phase 5.

The middle section of Table 1 displays the agreement after relabeling. We observe major improvement in agreement values for the most problematic cases shown in bold in the first section of the same table. The improvements, however, are not limited to the most problematic cases only, as we observe slight to moderate increase in all of the remaining agreement values as well.

4.8 Phase 8 (Tie-Breaking)

This phase included a meeting between the assessors, who went over all cases which could not be finalized by majority voting. Each assessor reconsidered their labels, having access to other assessors' labels. According to the bottom section in Table 1, the agreement values further increased after this phase, as expected. The duration of this phase was around three hours.

4.9 Phase 9 (Aggregation)

Each assessor passed their final labels to the coordinator, who then aggregated them via majority voting. As we will demonstrate in Section 6.3, resolving the ties considerably increased the percentage of questions which can be assigned a unique final label.

5 MULTI-FACETED QUESTION TAXONOMY

5.1 Question Intent

The `QuestionIntent` facet of our taxonomy contains the 16 intent categories presented below. We provide example questions for each intent category in Table 2.

- `Description`: This category covers mostly "what is X" or "what is X of Y" questions, where the user aims to obtain a definition or description of an object or one of its attributes.
- `Process`: These are typically "how to do X" questions that seek instructions, guidelines, or procedures which will facilitate an action to be performed by the user in real life.
- `Advice`: This category includes questions where the user aims to obtain personal advice on a particular topic. The `Advice` category differs from `Process` in that the former expects a somewhat subjective recommendation, whereas the latter expects an objective, step-by-step process description.
- `Opinion`: These are questions seeking to get a subjective opinion about a topic of interest (e.g., "what do you think about X" or "is X good/bad"). This class excludes `Advice` questions, where the subject is the user issuing the question.
- `Verification`: These are fact-checking questions that seek an affirmative yes/no answer which cannot be disputed.
- `Attribute`: "what is Y of X" questions that seek a particular property of a given named entity are in this category. We exclude questions whose answers are named entities to avoid a possible overlap with the `Quantity` and `Entity` categories.
- `Reason`: The expected answer to these questions include an explanation of causes underlying a certain action or event. Most questions of type "why is/do X Y" are in this category.
- `Location`: These are typically "where is X" questions that seek the position, address, or location of a given object or entity. The answers are not limited to geo-locations. For example, "where is X in human body" falls in this category.
- `Quantity`: These questions expect a numeric value as answer (e.g., price, frequency, duration, speed, age, length, weight).

Table 2: Question intent categories

Type	Examples
Advice	how can I be successful in life? how should I invest my salary?
Attribute	what is pristine edge's real name what is senegal's official language
Calculation	4,146.70+700+11900 1/2 cups in tbsp
Description	what is propylene kit what is oracle vpd functionality
Entity	who replaced ted kennedy in the senate who produced transformers
Language	what is puppy in swahili what is the common name for jade
List	types of aircraft southampton to guernsey types ant poison
Location	where are protists most abundant in humans what is oklahoma's absolute location
Opinion	is donald trump a good president? is ronaldo or messi a better player?
Process	what is needed to get home insurance how to check warranty of sd card sandisk
Quantity	how long is csus transfer orientation cost of an ice cream truck
Reason	why do knees swell up why do lipomas grow back
Resource	python temperature converter code tum mile love reprise lyrics english
Temporal	when do the oscar awards start when does daylight saving time return?
Verification	is tomorrow Monday? is donald trump the 34th president?
Weather	5 day weather forecast for york tybee island weather in march

- Entity: The expected answer to a question in this category is a named entity, excluding numeric entities.
- Language: These questions usually provide a name or description, and the expected answer is another name or object (e.g., "how is X called", "what is X in the Y language"). Questions seeking a named entity as answer are excluded.
- Temporal: These are typically the "when is X" questions, which expect to obtain a date or time of an event as answer.
- List: The expected answer is an itemized list whose items can have any type (e.g., entity, quantity, or a mix). For example, "which countries won the world cup" is a List question.
- Calculation: These are questions aiming to use the search engine as a calculator for arithmetic operations or unit conversion. Similar to the Quantity category, the expected answer is a numeric value, but the Calculation category contains one or more numeric values as input in the question.
- Weather: These are questions about weather forecasts.

Table 3: Named entity types

Type	Examples
Audio/Visual	Music albums, songs, movies
Event	Wars, concerts, competitions, ceremonies
Illness	Diseases, syndromes, conditions
Location	Countries, states, cities, towns, airports, addresses, building names, landmarks
Organization	Universities, companies, institutions, agencies, political parties
Person	People, mythological creatures, fictional characters
Product	Drugs, software, brands
Misc	Anything else (e.g., planets)
Date	July, 23/10/2019, 1956, Wednesday
Duration	2 hours, three years
Fraction	2/5, one-third
Money	$300, 1000 JPY
Number	1, 1.2, sixty
Ordinal	1st, 2nd, 3rd, fourth
Percent	5%, 43.3%
Range	5–20, [40, 50]
Set	Christmas day, Mother's day (recurring dates)
Time	10:30PM, 20:45, noon

Table 4: Question word types

Type	Description
What	Asking for information about something
When	Asking about time
Where	Asking in or at what place or position
Which	Asking about choice
Who	Asking what or which person (subject)
Whom	Asking what or which person (object)
Whose	Asking about ownership
Why	Asking for reason
How	Asking about manner
How + adj/adv	Asking about extent or degree
Is, are, do, does	Questions expecting a yes or no answer

- Resource: Non-informational questions where the goal is to obtain an online or offline resource are in this category, excluding questions with Calculation and Weather intents.

5.2 Other Facets

The remaining four facets of our taxonomy are as follows:

- QuestionEntityType: If the question does not mention any entity, it is labeled as NoEntity. If at least two different entity types are mentioned, it is labeled as MultipleEntityTypes. Otherwise, an entity type is selected from Table 3.
- AnswerEntityType: This facet is about the type of the entity expected in the answer of the question.
- QuestionWordType: If the question does not contain an explicit question word, it is labeled as NoQuestionWord. Otherwise, a question word type is chosen from Table 4.
- AnswerGranularity: Each question is associated with the ideal granularity of text that would form a concise answer to the question. The available categories are Phrase (one

Table 5: Agreement between assessors

Study Phase	Intent Clarity	Rose and Levinson	Facets: Question Intent	Question Entity Type	Answer Entity Type	Question Word Type	Answer Granularity	Avg. of all facets
Phase 5 (Labeling)	0.317	0.207	0.661	0.186	0.575	0.890	0.444	0.551
Phase 7 (Relabeling)	0.330	0.345	0.697	0.327	0.763	0.945	0.477	0.642
Phase 8 (Tie-Breaking)	0.410	0.397	0.719	0.388	0.769	0.948	0.493	0.663

Table 6: Change in the percentage of ties as the study progresses

Study Phase	Type of Ties	Intent Clarity	Rose and Levinson	Facets: Question Intent	Question Entity Type	Answer Entity Type	Question Word Type	Answer Granularity
Phase 5 (Labeling)	Tie ($M < 3$)	0.80	9.96	11.10	25.15	3.13	0.96	6.38
	No tie ($2 < M < 5$)	12.60	59.03	40.04	66.91	61.13	24.67	60.89
	No tie ($M = 5$)	86.60	31.02	48.86	7.94	35.74	74.37	32.73
Phase 7 (Relabeling)	Tie ($M < 3$)	0.80	8.29	9.64	15.50	2.13	0.59	6.04
	No tie ($2 < M < 5$)	12.00	34.61	35.03	70.18	33.25	13.96	56.45
	No tie ($M = 5$)	87.20	57.10	55.34	14.32	64.62	85.44	37.52
Phase 8 (Tie-Breaking)	Tie ($M < 3$)	0.40	1.68	3.68	6.34	1.47	0.00	3.73
	No tie ($2 < M < 5$)	11.80	39.75	39.85	79.07	35.29	14.82	59.39
	No tie ($M = 5$)	87.80	58.57	56.47	14.59	63.24	85.18	36.88

or more words), `List` (a list of phrases), `Sentence` (a well-formed sentence), `Passage` (a paragraph with multiple sentences), or `Document` (a page with multiple paragraphs).

6 RESULTS

6.1 Baselines

We compare various aspects of the created taxonomy (e.g., assessor agreement and distribution of questions in categories) with Rose and Levinson's taxonomy[31], frequently used in the IR literature. We do not have a strong baseline against which we can compare the proposed editorial study procedure. However, we note that most editorial studies in the literature follow a simple procedure where the assessors label items based on the provided guidelines in a single phase without further refinement of labels. Therefore, we believe that the first five phases of the procedure proposed herein forms an upper bound on the effectiveness of those simpler processes.

6.2 Editorial Agreement

Table 5 shows the agreement observed between the assessors after various phases of the editorial study. As we aimed for in the adopted procedure, the agreement values gradually increase after each study phase that involves reassessment of the labels. The average agreement, which was 0.551 after the initial labeling phase (Phase 5), increased to 0.642 after the relabeling phase (Phase 7) and to 0.663 after the tie-breaking phase (Phase 8). After Phase 5, a large increase is observed in agreement for the `QuestionEntityType` and `AnswerEntityType` facets, as some assessors reconsidered their interpretation of these facets and available entity types. We also observe a major improvement in agreement on Rose and Levinson's taxonomy after a particular assessor modified their labels.

The highest agreement is observed for the `QueryWordType` facet, which is relatively easy to assess. The agreement on our question intent categories `QuestionIntent` is higher than that on Rose and Levinson's categories. This indicates that our taxonomy is relatively less ambiguous to humans. We also observe that the assessors have somewhat higher agreement on the `AnswerEntityType` facet than on the `QuestionEntityType` facet. This may appear surprising at first, as one can expect the types of entities mentioned in the question to be more explicit to the assessors than the entity type in the potential answer of the question. However, as we will see later, a large fraction of expected answers are formed of long pieces of text, and thus are not entities. Labeling of such answers is relatively easy, explaining higher agreement values.

6.3 Tied Cases

The final labels are obtained via majority voting. However, ties may prevent attaining the majority for certain questions, in which case a label cannot be assigned to the question: Let A denote the number of assessors, and M denote the number of assessors who opted for the most popular label of a question. In our editorial study, $A = 5$. Therefore, if $M < 3$, a label cannot be assigned to a question since there is a tie among one or more labels. Otherwise, there is no tie, and the most popular label is assigned either by full agreement ($M = 5$) or partial agreement ($2 < M < 5$) between assessors.

In Table 6, we show the percentage of questions with a tie after a particular phase of the study. As mentioned in Section 6.1, the values attained after Phase 5 (Labeling) form an upper bound on those that can be attained by the simpler labeling procedures often adopted in the literature. As intended, the percentage of tied cases decreases as assessors refine their labels throughout the process. Even for facets with relatively low editorial agreement (e.g., `QuestionEntityType`), the percentage of tied cases remain below 7%, which lets us aggregate the labels via majority voting and assign a final label to a large fraction of questions.

6.4 Distribution of Final Labels

Out of 1000 questions, the IntentClarity of a small fraction of questions were assigned MultipleIntents or UnclearIntent labels (30 and 15 questions, respectively), and 4 questions were not assigned any label. Hence, the rest of the analysis is based on the remaining 951 questions that were labeled as SingleIntent.

Tables 7a and 7b display the distribution of intent categories for Rose and Levinson's taxonomy and our taxonomy, respectively. At first glance, Rose and Levinson's taxonomy seems to lead to a highly skewed label distribution as around 86.7% of the questions fall in only one category (Directed-Closed), while our intent categories result in a more even distribution of labels. According to Table 7b, the intent of 78.0% of the questions is categorized as Description, Quantity, or Entity. Intent categories that are potentially more difficult to be answered by a web search engine (e.g., Reason, Verification, Opinion) are relatively less common. We also observe some emerging categories of question intent, such as Calculation and Language, as modern web search engines provide new means to answer such questions more effectively. The rest of the analysis is carried out with 884 questions that could be assigned a label, excluding those labeled as Resource or Weather.

According to Table 7c about half of the questions mention at least one named entity. The most popular entity type is Audio/Visual, followed by Location and Illness. This may indicate a bias towards entertainment or health-related questions. Although the use of Location entities in questions is also very common, most of these questions do not actually seek a specific location, but rather use the given location as a contextual constraint for another type of information need (e.g., "what is the average salary in perth, wa").

According to Table 7d, around 47.3% of questions expect a named entity to be retrieved as the answer. The most common AnswerEntityType is Person, followed by the three entity types, Duration, Money, Number, which are usually associated with questions having Quantity intent. The Duration type usually stems from travel-related questions, where users ask about the time distance between two geo-locations, while the Money type frequently appears due to questions about the cost of various items or products.

Table 7e displays the distribution of question word types. The results are in line with those in previous tables. As an example, the number of Who questions is similar to the number of questions seeking a Person (see Table 7d). Surprisingly, we observe very few When and Where questions. This may be because web users prefer to access temporal and geo-location information by other means (e.g., online schedules, map applications), instead of search engines.

Finally, Table 7f shows the distribution of labels for the expected granularity of an answer. The reported result is striking: almost all questions can be answered with a single passage or a shorter piece of text, i.e., without presenting entire web documents to users. It is also interesting to note that both Phrase and Passage are more prominent granularity types than Sentence, which lies somewhere in between in terms of length.

Table 7: Distribution of final labels assigned to questions

(a) RoseAndLevinson

Label	#	%
Directed-Closed	824	0.867
Directed-Open	66	0.069
Obtain	21	0.022
Advice	14	0.015
Interact	4	0.004
Locate	2	0.002
List	1	0.001
Download	1	0.001
Entertainment	1	0.001
Navigational	1	0.001

(b) QuestionIntent

Label	#	%
Description	368	0.387
Quantity	238	0.250
Entity	136	0.143
List	43	0.045
Process	33	0.035
Resource	27	0.028
Calculation	19	0.020
Language	10	0.011
Reason	8	0.008
Attribute	8	0.008
Temporal	7	0.007
Location	6	0.007
Weather	5	0.005
Verification	5	0.005
Opinion	2	0.002
Advice	1	0.001

(c) QuestionEntityType

Label	#	%
NoEntity	523	0.592
Audio/Visual	69	0.078
Location	55	0.062
Illness	44	0.050
MultipleTypes	32	0.036
Person	28	0.032
Product	28	0.032
Organization	27	0.031
Number	10	0.011
Misc	6	0.007
Fraction	2	0.002
Time	1	0.001
Money	1	0.001
Date	1	0.001
Event	1	0.001

(d) AnswerEntityType

Label	#	%
No entity	452	0.511
Person	130	0.147
Duration	102	0.115
Money	79	0.089
Number	62	0.070
Location	8	0.009
Range	7	0.008
Organization	7	0.008
Percent	6	0.007
Illness	6	0.007
Date	6	0.007
Time	2	0.002
Misc	2	0.002
Ordinal	1	0.001
Set	1	0.001

(e) QuestionWordType

Label	#	%
What	365	0.413
NoQuestionWord	220	0.249
How+adj/adv	133	0.151
Who	129	0.146
How	22	0.025
Is/Are/Do/Does	5	0.006
When	4	0.005
Why	4	0.005
Where	1	0.001
Which	1	0.001

(f) AnswerGranularity

Label	#	%
Phrase	432	0.489
Passage	302	0.342
Sentence	69	0.078
List	47	0.053
Document	1	0.001

7 RELATED WORK

7.1 Intent Taxonomies

Broder [4] proposed the first taxonomy of Web Search (WS), classifying queries as navigational, informational, and transactional. He used a survey combined with inspection of a sample (400) of AltaVista search log queries. The taxonomy was extended by Rose and Levinson [31] where Broder's transactional category was replaced with a resource category and two sets of subcategories were formed. A modified version of Rose and Levinson's taxonomy was adopted by White et al. [40] to investigate the way search engines handle keyword vs natural language queries.

Table 8: Existing query intent taxonomies

Taxonomy	Year	Cat.	Domain	Data Source
Broder [4]	2002	3	WS	AltaVista
Rose and Levinson [31]	2004	12	WS	AltaVista
Li and Roth [20]	2006	6	IR	TREC
Gupta et al. [13]	2018	6	IR	TREC, SQuAD
Bu et al. [5]	2010	6	CQA	Baidu Zhidao
Chen et al. [6]	2012	3	CQA	Y! Answers
Liu et al.[22]	2008	4	CQA	Y! Answers
Liu and Jansen [23]	2015	3	SN	Twitter
Current work	2020	16	WS	Bing

Li and Roth [20] proposed a query intent taxonomy based on their analysis of 1000 topics taken from TREC. The six intents were abbreviation, description, entity, human, location, and numeric value. The taxonomy also had a large number of fine-grained categories focused on entity types. The taxonomy was later used for a question classification task [30]. Gupta et al. [13] augmented Li and Roth's taxonomy, targeting well-formed questions with the end goal of improving the performance of semantic question matching. Some categories in the taxonomy were combined while others were split in order to reduce the ambiguity. The taxonomy was developed using the Stanford Question Answering Dataset (SQuAD).

Question taxonomies have been built for community question answering (CQA). Liu et al. [22] extended Broder's taxonomy by adding a social category at the top level. The informational category was subdivided into constant and dynamic and dynamic was split into opinion, context-dependent, and open categories. Bu et al. [5] proposed another intent taxonomy, based on analysis of questions asked on Baidu Zhidao. The intent categories were fact, list, reason, solution, definition, and navigation. The taxonomy was used in a question classification task. Chen et al. [6] proposed a coarse-grained taxonomy, where the questions were classified into three categories: objective, subjective, or social. According to the labeling study conducted on a Yahoo Answers dataset, about 2% of the questions could not be classified to a specific category due to ambiguity. Liu and Jansen [23] focused on questions asked in social networks (SN). Their taxonomy placed questions into three categories: accuracy, social, and knowledge. A labeling task was carried out on 3000 questions sampled from Twitter.

In educational research, taxonomies are proposed for task-oriented tutorial dialogues [2], automatic question generation for tutoring and assessment [26], and student profiling [14]. There are query taxonomies in e-commerce [35, 43] and medicine [11]. A taxonomy specific to 'why' questions was proposed by Breja and Jain [3]. A taxonomy for temporal questions was proposed by Saquete et al. [33]. Pomerantz [29] describes a meta-taxonomy of general question taxonomies in literature.

The related work is summarized in Table 8. Our intent taxonomy focuses on questions issued in web search. We are not aware of any existing taxonomy which was specifically designed for questions asked in the context of web search. This limits the comparability of our taxonomy with existing taxonomies. Nevertheless, in our work, to give some intuition to the reader, we compared the proposed taxonomy with the taxonomy of Rose and Levinson [31], which is widely used in the IR literature.

7.2 Editorial Study Procedures

Human assessors were recruited in a wide range of areas to obtain editorial labels: retrieval evaluation [10, 12, 25, 38], query subtopic mining [16, 42], intent mining [18, 39], assessing email intent [32], distinguishing informational from non-informational content [27], user satisfaction [21], and image search evaluation [34, 41]. Sormunen [36] employed assessors to assess two topics from a small set of documents, and their judgments were compared for the sake of assessment practice. Cui et al. [9] created labels to test a query expansion technique. Disagreements in labels were resolved by discussion among three assessors. Ishikawa et al. [17] introduced a model to identify high-quality answers in CQA sites. To evaluate answer quality, four assessors were recruited. The kappa agreement between assessors was considered as part of the measurement. Verberne et al. [37] considered if searchers and external assessors classify the intent of queries in the same way. Noticeable differences were found between inter-assessor agreement and the agreement between assessors and searchers. The agreement measured between external assessors was also found not to be a good estimator of the validity of intent classifications.

8 CONCLUSION

We proposed a taxonomy for questions asked in web search engines. This taxonomy contained new intent categories, such as `Calculation`, `Language`, `Attribute`, and `Weather`, which were not present in earlier taxonomies proposed for web search queries. Despite being more fine-grained, the intent categories in our taxonomy were less ambiguous for human assessors compared to the those proposed by Rose and Levinson [31]. Also, the proposed taxonomy was shown to result in a more balanced category distribution, which may be important when training intent detection models.

Based on our taxonomy, we presented a picture of questions asked in web search and their expected answers. Our results indicated the emergence of new types of search intents, such as calculation/conversion, as search engines evolve and provide new functionality to users. We also observed that the volume of `Why`, `Where`, `When`, and `Is/Are/Do/Does` questions in our sample is quite small. It is not clear whether this is because search engines have relatively poor performance in answering these kinds of questions or because users have other means to satisfy their information needs (e.g., CQA sites for `Why` questions or map applications for `Where` questions). Finally, we found that retrieving a passage or a shorter piece of text is sufficient to properly answer almost all questions. This further motivates the ongoing transition from organic search results to direct answers in web search.

Despite the clear guidelines and extensive training, it was surprising to observe major inconsistencies in the labeling behaviour of some assessors for certain facets of the taxonomy. This observation indicated the usefulness of the iterative procedure we followed. We have demonstrated that aggregating labels without further refinement, as done in most editorial studies, leads to lower assessor agreement and hence may result in labels that are less reliable.

ACKNOWLEDGMENTS

This research was supported in part by the Australian Research Council (DP180102687).

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv:cs.CL/1611.09268

[2] Kristy Elizabeth Boyer, William J Lahti, Robert Phillips, MD Wallis, Mladen A Vouk, and James C Lester. 2009. An empirically-derived question taxonomy for task-oriented tutorial dialogue. In *Proceedings of the Second Workshop on Question Generation*. 9–16.

[3] Manvi Breja and Sanjay Kumar Jain. 2017. Why-type Question Classification in Question Answering System. In *FIRE (Working Notes)*. 149–153.

[4] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM, 3–10.

[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, 1119–1128.

[6] 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*. ACM, 823–828.

[7] Christian Stab Christian M. Meyer, Margot Mieskes and Iryna Gurevych. 2014. DKPro Agreement: An Open-Source Java Library for Measuring Inter-Rater Agreement. In *Proceedings of the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin, Ireland, 105–109.

[8] Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2012. Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing* 17, 4 (2012), 32–38.

[9] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. 2003. Query expansion by mining user logs. *IEEE transactions on knowledge and data engineering* 15, 4 (2003), 829–839.

[10] Thomas Demeester, Robin Aly, Djoerd Hiemstra, Dong Nguyen, and Chris Develder. 2016. Predicting relevance based on assessor disagreement: analysis and practical applications for search evaluation. *Information Retrieval Journal* 19, 3 (2016), 284–312.

[11] John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj* 321, 7258 (2000), 429–432.

[12] Riya Goswami, Somik Karmakar, Avik Bisai, and Alok Ranjan Pal. 2017. A knowledge based approach for long answer evaluation. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 773–777.

[13] 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*. 499–513.

[14] Fatima Harrak, François Bouchet, Vanda Luengo, and Pierre Gillois. 2018. Profiling students from their questions in a blended learning environment. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. ACM, 102–110.

[15] Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1, 1 (2007), 77–89. https://doi.org/10.1080/19312450709336664

[16] Yunhua Hu, Yanan Qian, Hang Li, Daxin Jiang, Jian Pei, and Qinghua Zheng. 2012. Mining query subtopics from search log data. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 305–314.

[17] Daisuke Ishikawa, Noriko Kando, and Tetsuya Sakai. 2011. What Makes a Good Answer in Community Question Answering? An Analysis of Assessors' Criteria.. In *EVIA@ NTCIR*. Citeseer.

[18] Bernard J Jansen and Danielle Booth. 2010. Classifying web queries by topic and user intent. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 4285–4290.

[19] Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th International Conference on Computational Linguistics*. Association for Computational Linguistics, 1–7.

[20] Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering* 12, 3 (2006), 229–249.

[21] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Satisfaction with Failure or Unsatisfied Success: Investigating the Relationship between Search Success and User Satisfaction. In *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 1533–1542.

[22] Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. 2008. Understanding and summarizing answers in community-based question answering services. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 497–504.

[23] Zhe Liu and Bernard J Jansen. 2015. A taxonomy for classifying questions asked in social question and answering. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1947–1952.

[24] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[25] Vanessa Murdock, Diane Kelly, W Bruce Croft, Nicholas J Belkin, and Xiaojun Yuan. 2007. Identifying and improving retrieval for procedural questions. *Information Processing & Management* 43, 1 (2007), 181–203.

[26] Rodney D Nielsen, Jason Buckingham, Gary Knoll, Ben Marsh, and Leysia Palen. 2008. A taxonomy of questions for question generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

[27] Daniel Palomera and Alejandro Figueroa. 2017. Leveraging linguistic traits and semi-supervised learning to single out informational content across how-to community question-answering archives. *Information Sciences* 381 (2017), 20–32.

[28] Bo Pang and Ravi Kumar. 2011. Search in the Lost Sense of "Query": Question Formulation in Web Search Queries and its Temporal Changes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 135–140. https://www.aclweb.org/anthology/P11-2024

[29] Jeffrey Pomerantz. 2005. A linguistic analysis of question taxonomies. *Journal of the American Society for Information Science and Technology* 56, 7 (2005), 715–728.

[30] Khyathi Chandu Raghavi, Manoj Kumar Chinnakotla, and Manish Shrivastava. 2015. Answer ka type kya he?: Learning to classify questions in code-mixed language. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 853–858.

[31] 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)*. ACM, New York, NY, USA, 13–19. https://doi.org/10.1145/988672.988675

[32] Maya Sappelli, Gabriella Pasi, Suzan Verberne, Maaike de Boer, and Wessel Kraaij. 2016. Assessing e-mail intent and tasks in e-mail messages. *Information Sciences* 358 (2016), 1–17.

[33] Estela Saquete, Patricio Martinez-Barco, Rafael Munoz, and Jose-Luis Vicedo. 2004. Splitting complex temporal questions for question answering systems. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 566.

[34] Yunqiu Shao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2019. On Annotation Methodologies for Image Search Evaluation. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), 29.

[35] Parikshit Sondhi, Mohit Sharma, Pranam Kolari, and ChengXiang Zhai. 2018. A taxonomy of queries for e-commerce search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 1245–1248.

[36] Eero Sormunen. 2002. Liberal relevance criteria of TREC-: Counting on negligible documents?. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 324–330.

[37] Suzan Verberne, Maarten van der Heijden, Max Hinne, Maya Sappelli, Saskia Koldijk, Eduard Hoenkamp, and Wessel Kraaij. 2013. Reliability and validity of query intent assessments. *Journal of the American Society for Information Science and Technology* 64, 11 (2013), 2224–2237.

[38] Ellen M Voorhees and Hoa Trang Dang. 2003. Overview of the TREC 2003 question answering track. In *Trec*, Vol. 2003. Citeseer, 54–68.

[39] Chieh-Jen Wang and Hsin-Hsi Chen. 2014. Intent mining in search query logs for automatic search script generation. *Knowledge and information systems* 39, 3 (2014), 513–542.

[40] Ryen W White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. queries in informational search tasks. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 135–136.

[41] Zhijing Wu, Ke Zhou, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Does Diversity Affect User Satisfaction in Image Search. *ACM Transactions on Information Systems (TOIS)* 37, 3 (2019), 35.

[42] Deng Yi, Yin Zhang, and Baogang Wei. 2016. Query Subtopic Mining via Subtractive Initialization of Non-negative Sparse Latent Semantic Analysis. *J. Inf. Sci. Eng.* 32, 5 (2016), 1161–1181.

[43] Qian Yu and Wai Lam. 2018. Product Question Intent Detection using Indicative Clause Attention and Adversarial Learning. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 75–82.