ORCAS-I: Queries Annotated with Intent using Weak Supervision

Daria Alexander
Radboud University & Spinque
Utrecht, The Netherlands
daria.alexander@ru.nl

Wojciech Kusa
TU Wien
Vienna, Austria
wojciech.kusa@tuwien.ac.at

Arjen P. de Vries
Radboud University
Nijmegen, The Netherlands
arjen@cs.ru.nl

ABSTRACT

User intent classification is an important task in information retrieval. In this work, we introduce a revised taxonomy of user intent. We take the widely used differentiation between navigational, transactional and informational queries as a starting point, and identify three different sub-classes for the informational queries: instrumental, factual and abstain. The resulting classification of user queries is more fine-grained, reaches a high level of consistency between annotators, and can serve as the basis for an effective automatic classification process. The newly introduced categories help distinguish between types of queries that a retrieval system could act upon, for example by prioritizing different types of results in the ranking.

We have used a weak supervision approach based on Snorkel to annotate the ORCAS dataset according to our new user intent taxonomy, utilising established heuristics and keywords to construct rules for the prediction of the intent category. We then present a series of experiments with a variety of machine learning models, using the labels from the weak supervision stage as training data, but find that the results produced by Snorkel are not outperformed by these competing approaches and can be considered state-of-the-art. The advantage of a rule-based approach like Snorkel's is its efficient deployment in an actual system, where intent classification would be executed for every query issued.

The resource released with this paper is the ORCAS-I dataset: a labelled version of the ORCAS click-based dataset of Web queries, which provides 18 million connections to 10 million distinct queries. We anticipate the usage of this resource in a scenario where the retrieval system would change its internal workings and search user interface to match the type of information request. For example, a navigational query could trigger just a short result list; and, for instrumental intent the system could rank tutorials and instructions higher than for other types of queries.

CCS CONCEPTS

• **Information systems** → **Query intent**; *Web log analysis*; • **Computing methodologies** → *Semi-supervised learning settings.*

KEYWORDS

intent labelling, weak supervision, click data, Snorkel, web search

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531737

ACM Reference Format:
Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. ORCAS-I: Queries Annotated with Intent using Weak Supervision. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477495.3531737

1 INTRODUCTION

When a user types a query into a search engine, there is usually a specific intent behind it: to download something, to purchase a product, to find a particular site or explore a topic. Understanding that intent can be very useful for providing relevant results to the searcher and increasing the value of the information obtained. Also, tailoring the content of the site to user intent helps increase the site's visibility rates.

While manual classification of user intent provides more accurate labels, manual annotation of large amounts of data can be very challenging. A weak supervision approach allows avoiding hand labelling of large datasets, which can save a lot of time and energy. In weak supervision, noisy labels are generated by using heuristics in the form of domain-specific rules or by using pattern matching. In this paper, we aim to understand how to automatically label click log data with user intent and annotate a large click dataset with those labels using weak supervision.

Commercial Web search engines refrain from disseminating detailed user search histories, as they may contain sensitive and personally identifiable information [1]. In the past, datasets such as AOL revealed personal information about the users, for example, their location and their names [6]. ORCAS [12], a new dataset released by Microsoft deals with that issue by not providing anything that could potentially help to identify the searcher. The absence of personal information and the large size of this dataset makes it very interesting for researchers, yet also makes impossible to analyze aspects like user behaviour during a search session.

Although many studies performed an automatic classification of user intent in search log data [5, 18, 21, 23, 28, 30], there were fewer papers addressing this subject recently [15, 34]. Also, to our knowledge there are no released large click datasets labelled with user intent. The datasets where user intent is annotated are mainly used for task-oriented dialogue systems. For example, MANtIS, a large-scale conversational search dataset containing more than 80,000 conversations across 14 domains that are englobing complex information needs [39]. Another dataset [27] was collected via crowd-sourcing and consists of 23,700 utterances covering 150 different intents. The Schema-Guided Dialogue dataset [40] has over 16,000 dialogues in the training set belonging to 16 domains. The intents in those datasets often differ from the intents for search log queries and are specific to interactions with conversational agents, such as "transfer", "make payment" and "to do list update" [27, 40].

To fill this gap, we suggest using recent labelling techniques such as weak supervision combined with methods previously employed for intent classification of search log queries, such as a rule-based approach. We propose a user intent taxonomy based on a taxonomy established by Broder [8] that divides intents into three levels: informational, navigational and transactional. We perform the classification on two levels: 1) three categories of Broder's taxonomy, 2) three subcategories in the informational class: factual, instrumental and abstain. We base our automatic classification on Jansen et al.'s [18] user intent characteristics, upon which we improve to further increase the quality of the taxonomy. Then we perform the labelling of the ORCAS dataset, which has 18 million connections to 10 million distinct queries. For the labelling, we use weak supervision with Snorkel [41]. After that, we train five different machine learning models on the 2 million items subset of the ORCAS dataset.

Our main findings are as follows:

- Our automatic labelling provides better results than those reported in the original study;
- classifying the queries on the three top level categories provides better scores than classifying them on the full taxonomy;
- benchmark models do not significantly outperform the Snorkel classifier;
- the lack of performance of the models can be explained by 1) the specificities of weak supervision, 2) the lack of external knowledge such as ontologies, 3) the length of the queries and the absence of grammatical structure in them.

This work makes the following contributions:

- We improve the existing characteristics for automatic intent classification of search log queries;
- we suggest subcategories that allow to have a more fine-grained automatic classification of user intent;
- we share a publically available annotation for the widely used ORCAS dataset.

2 RELATED WORK

The related work relevant to this paper is linked to intent labelling, automatic classification of user intent and weak supervision.

2.1 Intent labelling

When users type queries in search engines, they often have a specific intent in mind. Broder [8] divides queries into three categories according to their intent: navigational, transactional and informational. An informational intent refers to acquiring some information from a website, a navigational intent consists of searching for a particular website, a transactional intent refers to obtaining some services from a website (e.g. downloading the game). In Broder's study, queries from AltaVista query log were classified manually and information about clicked URLs was not used. This taxonomy was expanded in [43] with sub-classes for informational, navigational and transactional categories. Contrary to Broder, clicked URLs were used for intent classification, but did not show significant improvement compared to labelling of queries only. The following studies used the complete Broder's taxonomy [18, 23] or some of its categories [17, 21, 28–30]. Some studies added other categories, such as *browsing* [24] or *learn* and *play* [44].

2.2 Automatic classification of user intent

Early studies that performed automatic classification of user intent were usually limited to only two of Broder's categories: either informational and navigational [21, 28], or informational and transactional [5]. They adopted different techniques such as computing the scores of distribution of query terms [21], classification of queries into topics [5] as well as tracking past user-click behavior and anchor-link distribution [28].

In order to automatically classify search intent, researchers used click features. They found that if the intent of a query was navigational, then users mostly clicked on a single website. On the other hand, if the intent was informational, users clicked on many results related to the query [28, 30]. URL features, which take into account the text of the URLs were considered important for navigational category along with click features [32]. Also, using the text of the clicked URLs improved the results for the navigational category but not for the informational category [21].

Jansen et al. [18] established a rule-based approach and defined query characteristics for automatic classification of informational, navigational and transactional intents. They were linked to query length, specific words and combinations of words encountered in queries, and to the information about the search session (e.g. whether it was the first query submitted). Assigning labels according to the established characteristics was done as a first step before using machine learning approaches, such as performing k-means clustering [23]. In order to add some additional features to Jansen et al.'s characteristics, natural language processing techniques such as POS-tagging [15, 34], named entity recognition and dependency parsing [15] were used, however, the classification was done on much smaller datasets than in the original study.

2.3 Weak supervision

One of the most common problems with successful training of machine learning models is the lack of datasets with good quality annotations. Manual collection of annotations is a costly, tedious and time-consuming process. Academic research institutions often do not have enough funding to gather large-scale annotations, limiting their capabilities of creating significant corpora. This became even more visible with the growth of large pre-trained language models [9, 14] that led to impressive gains on many natural language understanding benchmarks [46], requiring a large number of labelled examples to obtain state-of-the art performance on downstream tasks [49]. In order to mitigate the lack of labelled data, recent works tried using other approaches to produce annotated datasets, like the usage of regular expression patterns [3] or class-indicative keywords [22].

Weak supervision is an approach in machine learning where noisy, limited, or imprecise sources are used instead of (or along with) gold labelled data. It became popularised with the introduction of the data programming paradigm [42]. This paradigm enables the quick and easy creation of labelling functions by which users express weak supervision strategies or domain heuristics. Various weak supervision approaches can be represented by labelling functions, such as distant supervision, heuristics or the results of crowd-sourcing annotations. Weak supervision has already been

successfully applied in other problems in the area of natural language processing and information retrieval [4, 13, 16]. In this paper, we focus on the usage of heuristics to create the labelling functions for intent classification.

Snorkel is a weak supervision system that enables users to train models using labelling functions without hand labelling any data [41]. It is an end-to-end system for creating labelling functions and training and evaluating the labelling model. It is designed to work with classification or extraction tasks. According to the recent benchmark study [49], it still offers comparable performance to newer and more complex weak supervision systems.

3 TAXONOMY

Researchers do not agree on the terms to use for search behaviour classification. Some researchers use the notion of *intent* for determining informational, navigational and transactional search behaviour [8, 18, 23]. Others use the term *goal* instead of the term *intent* for the same taxonomy [43, 44]. There is also a group of researchers who use the term *task* for the taxonomy that contains all [45] or some [24] of Broder's categories.

A search task is defined as a task that users need to accomplish through effective gathering of information from one or several sources [10, 31]. A task can have a goal, which is a part of a task description. A search intent is defined as the affective, cognitive, or situational goal as expressed in an interaction with information systems. [18].

As a search intent is an expression of a goal in an interaction with information systems and we analyse the data that reflects this interaction, we decided to adopt Broder's choice of terms and use the notion of *intent*. For our study, we use Broder's initial intent classes: *informational*, *navigational* and *transactional*. We refine the informational class with three subcategories: 1) *factual* [20, 24], 2) *instrumental* (also called *advice* in [43] or *learn* in [44]) and 3) *abstain*.

The *Factual* and *instrumental* subcategories were chosen because it was possible to identify characteristics that would allow their automatic classification and potentially many queries exist that would have those intents. We also considered other subcategories such as *descriptive* [25] and *locate* intents [43], but found that they were to narrow for our goal of allowing classification. We provide the taxonomy (Figure 1) as well as categories definitions.

- **Navigational intent**: the immediate intent is to reach a particular website [8];
- **Transactional intent**: locate a website with the goal to obtain some other product, which may require executing some Web service on that website [18];
- **Informational intent**: locate content concerning a particular topic in order to address an information need of the searcher [18];
  - **Factual intent**: locate specific facts or pieces of information [24];
  - **Instrumental intent**: the aim is to find out what to do or how to do something [25];
  - **Abstain**: everything inside the informational category that is not classified as factual or instrumental.

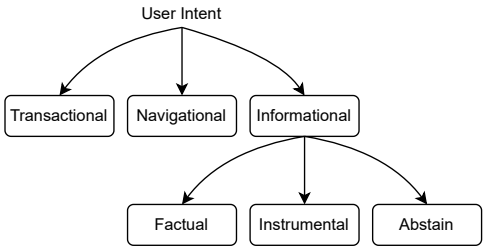

Figure 1: User intent taxonomy used in this study.

4 DATASET

To classify user intent of search queries according to Jansen's characteristics, previous studies used the Dogpile transaction log [19] or the AOL web query collection [37]. Both of those datasets contained user IDs, which led to some privacy issues.

Concerned with those privacy problems, we decided to perform automatic annotation on a dataset that does not have user IDs. The ORCAS dataset appealed to us for 3 main reasons: 1) it does not contain information about users 2) it is a contemporary large dataset 3) it contains general queries such as one can find in any search engine.

The ORCAS dataset is part of the MS MARCO datasets (Microsoft) and is intended for non-commercial research purposes. It contains 18.8 million clicked query-URL pairs and 10.4 million distinct queries. The dataset has the following information: *query ID*, *query*, *document ID* and *clicked URL*. The documents that the URLs lead to come from TREC Deep Learning Track.

This dataset was aggregated based on a subsample of Bing's 26-month logs to January 2020. The creators of the dataset applied several filters to the log. Firstly, they only kept query-URL pairs where the URL is present in the 3.2 million document TREC Deep Learning corpus. Secondly, they applied a k-anonymity filter and only kept queries that are used by k different users. Finally, offensive queries such as hatred and pornography were removed.

For labelling, we use a 2-million sample of the ORCAS dataset. This is a sample that is chosen randomly from the whole dataset. We call this dataset **ORCAS-I-2M**. As the dataset is already preprocessed, we did not need to do any additional pre-processing. For example, the text of the queries is already lower cased. We decided to keep the punctuation in the queries because when the user is searching for a specific site, the dots before the domain names (such as ".com", ".org") can help to assign the right label to those queries. We also decided to keep multiple instances of the same query because they can potentially have a different label, depending on the contents of the URL clicked.

5 METHODOLOGY

In this section, we describe the process of creating the characteristics of user intent provided in our taxonomy. Those characteristics enable the automatic classification of the intent.

5.1 Establishing characteristics for each label

The automatic assignment of labels to queries was based on the characteristics established by Jansen et al.[18] for transactional,

navigational and informational intents. However, we re-evaluated the characteristics for transactional and navigational categories and suggested new ones. Also, as we defined two subcategories (factual and instrumental) inside informational category, we decided to reuse some of Jansen et al.'s characteristics for this class and add new ones for each subcategory.

To determine user intent characteristics, queries drawn from different datasets (AOL web query collection [37], TREC 2014 Session Track [11], MS MARCO Question Answering Dataset [36]) were analysed. For each characteristic we annotated small subsets of the datasets automatically and then for the next subsets we adjusted the characteristics of the classification; the characteristics that did not improve automatic labelling, for example, when they did not cover a significant part of the data, were discarded.

A big difference between our classification and Jansen's classification is that we do not only use queries but also URLs, as suggested by [32] and [21]. That helped us to refine the classification and include more features that could help to assign a label to a query.

We present the characteristics that we took from Jansen et al. as they were, those that we changed and those that we created ourselves. They are presented by category.

5.1.1 *Transactional intent.*

We kept the majority of the characteristics from the transactional category. They are linked to various keywords that one would use when performing a transaction, such as "download", "buy", "software".
The characteristics we kept are:

- queries with "download" terms (e.g. "download", "software");
- queries relating to image, audio, or video collections;
- queries with "audio", "images", or "video" as the source;
- queries with "entertainment" terms (pictures, games);
- queries with "interact" terms (e.g. "buy", "chat");

The ones we did not use:

- queries with "obtaining" terms (e.g. lyrics, recipes);
- queries containing terms related to movies, songs, lyrics, recipes, images, humor, and pornography;
- queries with compression file extensions (jpeg, zip).

As for the characteristics that we did not take, we empirically understood that 1) many queries that contain movies, songs, lyrics and recipes terms belong to the factual subcategory and 2) extensions such as "jpeg" and "zip" do not usually indicate transactional intent. For example, "zip" usually appears in phrase "zip code", which would belong to factual category. Also, many queries that contain the term "jpeg" would be classified under the instrumental category, such as "converting to jpeg".

5.1.2 *Navigational intent.*

We kept two of Jansen et al. characteristics for navigational intent. For the queries containing domain names we used a list of top-level domains that we crawled from the Wikipedia[1].

[1]https://en.wikipedia.org/wiki/List_of_Internet_top-level_domains

The characteristics we kept are:

- queries containing domains suffixes;
- queries with "Web" as the source;

The ones we did not use:

- searcher viewing the first search engine results page;
- queries length (i.e., number of terms in query) less than 3;

The ones that we refined:

- queries containing company/business/organization/people names;
- our version: queries for which the Levenshtein ratio between the query and the domain name is equal or greater than 0.55;

We did not take the characteristic of "searcher viewing the first results page", because, unlike Jansen et al., we do not have access to information about user sessions. We also decided that considering queries shorter than 3 words as navigational is counter-productive as 35% of queries in the data have fewer than 3 words so it can potentially lead to many false positives. For example the queries "allergic rhinitis" and "generation terms", despite having only two words, do not belong to navigational category.

As for the "queries containing company/organization/people names", we found that navigational queries do not only contain the names of organizations and people. For example, the query "army study guide" leads to the site *https://www.armystudyguide.com* which is dedicated to US army study guide. Instead, we identified navigational intent by considering the similarities between the queries and the domain name parts of URLs. We used the Levenshtein similarity ratio which is calculated according to the following formula:

$$\frac{|a| + |b| + \mathrm{Lev}(a, b)}{|a| + |b|}$$

Here, $\mathrm{Lev}(a, b)$ is Levenshtein distance (the minimum number of edits that you need to do to change a one-word sequence into the other) and $|a|$ and $|b|$ are lengths of sequence a and sequence b respectively. A threshold on Levenshtein ratio was empirically established at 0.55, which means that if the query and the domain name were 55% or more similar they were classified as navigational.

5.1.3 *Informational intent.*

In his study, Jansen classified 81% of queries as having informational intent. Follow-up research considered this category too broad [44]. It was one of the reasons that motivated us to introduce subcategories inside the informational category: factual, instrumental and abstain.

Factual intent. Jansen et al.'s characteristics for informational intent such as having "queries with natural language terms" and "queries length greater than 2" are too broad to be useful. We did use "queries that contain question words" though. We suggest the following characteristics for factual intent:

- queries containing question words (e.g. "what", "when", "where", "which");
- queries starting with question words (e.g. "can", "does");
- queries containing words such as "facts", "statistics" and "quantities";

Table 1: Most popular sites that provide specific facts in the 2M sample of ORCAS dataset.

Name of the site	Count
wikipedia.org	287,269
webmd.com	23,110
merriam-webster.com	19,881
drugs.com	14,177
dictionary.com	9,501
mayoclinic.com	8,670
reference.com	8,670
britannica.com	7,894
medicinenet.com	7,136
accuweather.com	6,041
weather.com	5,893

- queries containing the terms linked to cost or price (e.g. "average", "cost", "amount", "sum", "pay");
- queries containing words that can be replaced by numbers (e.g. "phone", "code", "zip");
- queries containing words of definition (e.g. "define", "definition", "meaning");
- the clicked URL leads to the sites that contain specific facts.

Usually, queries that contain question words require specific answers, "what's the fastest animal in the world", so the intent here is to find facts. After analysing search queries in different datasets, we realised that queries that contain words associated with quantities, price and money have factual intent. Also, people searching for a term or concept definition usually look for specific information.

In order to get sites that provide specific facts, we took the 20 most frequent sites in the dataset and identified these sites among them (see Table 1). Usually those are encyclopedias or sites that contain some specific information (such as information about drugs, local weather etc.).

Instrumental intent. No characteristics in Jansen et al. are relevant to instrumental intent, except "queries that contain question words". For the queries that are aimed at finding resources about what to do or how to do it, we established the following characteristics:

- queries containing question words (e.g. "how to", "how do", "how does");
- queries that begin with infinitive form of a verb (e.g. "build", "cook");
- queries that begin with the "-ing" form of the verb (e.g. "making", "doing");
- the clicked URL leads to the sites that contain tutorials and instructions.

We figured out that the queries issued for finding advice or instructions often start with "how to" and "how do". Also, queries that start with a verb in the infinitive or using the "-ing" form usually have instrumental intent. For identifying the infinitives

Table 2: Most popular sites that provide instructions and tutorials in the 2M sample of ORCAS dataset.

Name of the site	Count
support.office.com	9,641
support.apple.com	7,494
wikihow.com	5,307
support.google.com	4,348

of the verbs we used a list of 850+ common English verbs[2]. For queries that begin with the "-ing" form of a verb we chose *Spacy* [3] to determine whether the part-of-speech of the first word is a verb and whether it has an "-ing" postfix. As well as for factual queries, we used the clicks to the sites among the 20 most popular sites, in this case to those that provide tutorials and advice.

Abstain category. The queries that were not classified as transactional, navigational, factual or instrumental are assigned to abstain category. According to Jansen et al., the queries that do not meet criteria for navigational or transactional have informational intent. Thus, we decided to make abstain subcategory part of the informational category. However, we could not establish consistent automatic characteristics for this group of queries because we could not find any reliable patterns in them.

What are those abstain queries? We expect that many of these would belong to an *exploratory* intent, when a user wants to learn or investigate something, but the goal of this search is amorphous [20, 33]. Having user sessions usually helps to understand if the queries are exploratory. An exploratory search process is described as submitting a tentative query, then exploring the retrieved information, selectively seeking and passively obtaining cues about where the next steps lie [47]. As we do not have user sessions, establishing characteristics for those queries is infeasible for now.

5.2 The test dataset

In order to evaluate the performance of our weak supervision approach, we manually created a test set collection. We randomly selected 1000 queries from the original ORCAS dataset that were not in the ORCAS-I-2M dataset. The test set was annotated by two IR specialists using the open-source annotation tool *Doccano*[4]. In case of doubt about assigning a specific intent to a query, the result pages of clicked URLs were used as an additional hint for classification. For example, if the query intent was unclear and the result page was a tutorial, the query was classified as having instrumental intent. For inter-annotator agreement on the test set, the Cohen Kappa statistic was 0.82. Remaining disagreements between the two annotators were then resolved by discussion, leading to a final decision for every query. We call this manually annotated dataset **ORCAS-I-gold**.

[2]https://github.com/ProjectDossier/intents_labelling/blob/main/data/helpers/verbs.txt
[3]https://spacy.io/
[4]https://github.com/doccano/doccano

5.3 Creating Snorkel labelling functions

In machine learning terms, our intent taxonomy could be represented as a two-level, multi-class classification problem. Snorkel has originally only been implemented to handle annotations for single-level classification problems. As our taxonomy is hierarchical, we needed to define two layers of Snorkel labelling functions.

We defined the first level of labelling functions to distinguish between navigational and transactional intents. All the queries that could not fit into one of these two categories were classified as informational intent in our taxonomy. Based on the characteristics defined in Section 5.1, we created four labelling functions for navigational queries and three functions for transactional queries.

On the second level, we defined labelling functions to cover factual and instrumental intents. Similar to the previous step, we designed nine factual functions and four instrumental labelling functions, using the characteristics from Section 5.1. All queries that were not assigned a label from the two layers of Snorkel got an abstain category.

We initially used *Spacy*'s *en_core_web_lg* language model to identify part of speech information and to detect named entities. After initial analysis, this proved to generate too many false negatives, especially for the detection of verbs. For example, the queries "change display to two monitors" and "export itunes library" were misclassified as abstain, because the verbs "change" and "export" were labelled as nouns. We suspect that these errors were primarily caused by the lack of a proper sentence structure, which prevented *Spacy* from correctly detecting the part of speech of the word. In the final version, we decided to use a list of the 850+ common verbs with which we obtained comparable coverage with fewer false positives. Eventually, we have only used *Spacy* for a labelling function where queries begin with the "-ing" form of the verb.

5.4 Training Snorkel

To obtain a final prediction score, we run independently two levels of Snorkel annotations. Based on our classification that all non-transactional, non-navigational queries are informational, for the second level prediction, we use all the queries which were assigned abstain from the first level. In order to conduct label aggregation, we experiment with both the LabelModel and MajorityLabelVoter methods implemented in Snorkel. LabelModel estimates rule weights using an unsupervised agreement-based objective. MajorityLabelVoter creates labels by aggregating the predictions from multiple weak rules via majority voting. We test their predictions on the test dataset using default hyperparameters. Results are presented in Table 3.

Table 3: Comparison of Snorkel labelling models results on ORCAS-I-gold.

Model	Metric	Precision	Recall	F1-score
Majority Label Voter	Macro avg	.780	.763	.771
	Weighted avg	.786	.783	.783
Label Model	Macro avg	.737	.773	.750
	Weighted avg	.779	.770	.772

After analysis of results on the testset, MajorityLabelVoter achieved higher scores for all measures except for macro average recall. Therefore, we decided to use it to obtain the final labels for the ORCAS-I-2M dataset. MajorityLabelVoter has the additional benefit that it provides more explainable results, as for every query, the user can be presented with the raw aggregation of the single labelling functions.

6 BENCHMARK MODELS

We benchmark five different models by training them on the ORCAS-I-2M dataset and evaluating the results on ORCAS-I-gold. We split the ORCAS-I-2M dataset into train and validation sets with 80:20 ratio. Hyperparameters not mentioned below are given their default values.

- **Logistic regression**: For logistic regression we use tf-idf for text representation and the sklearn [38] standard scaler for feature scaling.
- **SVM**: Support Vector Machine. As our dataset is large, we use linear support vector classification. As for logistic regression, we use tf-idf vector for text representation.
- **fastText**: We use the fastText [7] model with word embeddings trained from scratch on ORCAS-I-2M dataset. This is a text classification model that uses average of word embeddings to compute the representation of a document. We utilise the Python wrapper implementation[5].
- **BERT**: We use pre-trained, 110M parameters BERT model [14] followed by a classification head. We use the *bert-base-uncased* model implemented in the HuggingFace library [48]. Batch size is set to 64, fine-tuning is conducted for 10 epochs.
- **xtremedistil**: We also evaluate *xtremedistil-l6-h384-uncased* checkpoint from the XtremeDistilTransformers model [35]. Similar to BERT, we fine-tune it for 10 epochs with a batch size set to 64.

7 RESULTS

In this section, we present the statistics of ORCAS-I annotated both manually and with Snorkel. We also show the results of training the benchmark models on ORCAS-I-2M when evaluated on ORCAS-I-gold.

7.1 Snorkel classification

7.1.1 Top level categories. We run the Snorkel classifier on the ORCAS-I-gold testset. Table 5 shows that it attains an F1-score of 0.82 and an accuracy of 0.90. The category with the best performance is informational, followed by transactional and navigational.

Table 4 shows how our approach outperforms the results of the original Jansen et al. paper and also those reported in other studies, except [23], that reaches an accuracy of 0.94. However, the informational category is over-represented in the ORCAS-I-gold, as well as in ORCAS-I-2M dataset (see section 7.4 and Table 10), which can potentially influence the quality of training of the models such as BERT.

7.1.2 Full taxonomy. Table 6 shows that in the full intent taxonomy, the factual subcategory has the best results. It is followed

[5]https://pypi.org/project/fasttext/

Table 4: Accuracy of Snorkel classifier compared to other studies.

Study	Dataset	# queries in the dataset	Features	Algorithm	Accuracy
Jansen et al. [18]	Dogpile transaction log	4,056,374	queries only	rules	74%
Ashkan et al. [2]	data from Microsoft adCenter	800,000	queries only	SVM	86%
Kathuria et al. [23]	Dogpile transaction log	4,056,374	queries only	k-means	94%
Figueroa [15]	data from the AOL query collection	4,811,638	queries and URLs	MaxEnt	82.22%
Our study	ORCAS	18,823,602	queries and URLs	rules	90.2%

Table 5: Detailed Snorkel weak labelling results for the test set using only three top level categories on ORCAS-I-gold.

Category	Precision	Recall	F1-score	Examples
Navigational	.776	.731	.753	171
Transactional	.756	.791	.773	43
Informational	.936	.945	.941	786
Macro Avg	.823	.822	.822	1000
Weighted avg	.901	.902	.901	1000
	Accuracy		.902	1000

Table 6: Detailed results for Snorkel weak labelling for full intent taxonomy on ORCAS-I-gold.

Model	Precision	Recall	F1-score	Examples
Navigational	.800	.725	.761	171
Transactional	.756	.791	.773	43
Instrumental	.774	.695	.732	59
Factual	.847	.826	.837	363
Abstain	.723	.780	.750	364
Macro avg	.780	.763	.771	1000
Weighted avg	.786	.783	.783	1000
	Accuracy		.783	1000

by navigational category, which gets slightly higher precision results compared to predictions from three top level categories. The transactional and abstain categories as well as the instrumental subcategory perform worse than the others. For the transactional and instrumental categories this result can be linked to the small number of queries of this type in ORCAS-I-gold.

The results show that using more categories lowers the overall F1-score and accuracy. We can conclude that having more classes and more rules can potentially diminish Snorkel performance. Nevertheless, having subcategories in the informational category would allow to provide more choice for distinguishing between different types of intent.

7.2 Benchmark models

To train the models on ORCAS-I-2M, we use two types of training data: just the query and query plus URL. URL features help improve

Table 7: Macro average scores comparison for all benchmark models trained on three top level categories. Underlined scores indicate the highest score within the different input features for each model, bold values indicate the highest score overall.

Model	Input features	Precision	Recall	F1-score
Snorkel	query	**.864**	.700	.742
	query + URL	.822	.822	.822
Logistic regression	query	.796	.756	.770
	query + URL	.815	.758	.784
SVM	query	.842	.783	.798
	query + URL	.824	.817	.817
fastText	query	.788	.737	.748
	query + URL	.820	.801	.806
BERT	query	.821	.785	.796
	query + URL	.832	**.823**	**.826**
xtremedistil	query	.846	.771	.790
	query + URL	.818	.818	.817

Table 8: Macro average scores comparison for all benchmark models trained on all the categories. Underlined scores indicate the highest score within the different input features for each model, bold values indicate the highest score overall.

Model	Input features	Precision	Recall	F1-score
Snorkel	query	.771	.648	.667
	query + URL	.779	.764	.770
Logistic regression	query	.701	.611	.643
	query + URL	.714	.689	.700
SVM	query	.735	.689	.703
	query + URL	.782	.759	.767
fastText	query	.694	.643	.660
	query + URL	.768	.753	.758
BERT	query	.742	.705	.717
	query + URL	**.789**	.764	**.774**
xtremedistil	query	.725	.691	.696
	query + URL	.781	**.765**	.772

classification effectiveness. Tables 7 and 8 show that when we eliminate URL features from Snorkel (we mute or change the labelling functions that are using URLs) especially recall is reduced. This is particularly noticeable for the navigational category, for which recall drops from 0.73 to 0.35. This confirms the findings of [21] that using the text of clicked URL improves the results for navigational category.

We hypothesise that as we take URL features into account for the Snorkel classifier, models that train on queries and URLs will outperform the models that train on queries only. This hypothesis is confirmed for the full taxonomy, especially for fastText and xtremdistil. For the three top level categories, the difference in performance is not so large (except for fastText), which can be explained by fewer URL features for these categories.

Even if we only use the query, the recall for the models trained on the three top level categories is higher than the recall of Snorkel without URL functions. As for the full taxonomy, SVM, BERT and xtremdistil show improvements on recall for query-only when compared to Snorkel. It indicates that the models learn well from the labels assigned by the Snorkel query and URL functions, even if they are trained on queries only.

None of our benchmark models significantly outperforms our Snorkel baseline when trained on queries and URLs. This could have been an expected behaviour when comparing two models, one being the teacher and the other the student who learned only from this one teacher, without any external knowledge. We also hypothesise that transformer-based models cannot express their full power because the input sequences are, on average, very short and often do not have a proper grammatical structure.

While our experiment does not indicate how the machine learning models could strictly outperform the base, weak labelling, we see some potential directions for future work. One solution would be to use more than one annotated click log dataset, ideally using different annotation types (i.e. a either combination of weak and human annotations or two distinct weak supervision models). Another solution would be to use a model that could utilise an external knowledge base or ontology to understand the nuances between different categories, which often depend on the type of website that the user selected.

7.3 Final intent classification

After analysing results on ORCAS-I-2M for both Snorkel and other benchmark models we decided to use the Snorkel model to predict intent categories on the full ORCAS dataset. We call this dataset **ORCAS-I-18M**. As ORCAS-I-18M contains all items that were included in both ORCAS-I-gold and ORCAS-I-2M, it should be used with caution when training and evaluating machine learning models.

7.4 Dataset statistics

Overview statistics of all ORCAS-I datasets used in this study are presented in Table 9. ORCAS-I-2M covers more than half of the unique domains available in ORCAS-I-18M, and at the same time, more than 40% of unique URLs. This means that even though ORCAS-I-2M constitutes only around 10% of the ORCAS-I-18M elements; it is still a representative sample. The mean length of the query is comparable in all ORCAS-I datasets. We noticed that 246 duplicated query-URL pairs exist in the ORCAS-I-18M dataset and 7 in ORCAS-I-2M. Even though we do not preserve uniqueness in our training data, such a small amount of duplicates would not affect the training of machine learning models.

Table 9: Statistics of ORCAS datasets used in this paper ("un." stands for "unique").

	ORCAS-I-gold	ORCAS-I-2M	ORCAS-I-18M
dataset size	1,000	2,000,000	18,823,602
un. queries	1,000	1,796,652	10,405,339
un. URLs	995	618,679	1,422,029
un. domains	700	126,001	241,199
un. words in query	2,005	334,724	1,380,160
mean query length (words)	3.21	3.25	3.25

We measure the label distribution for all three annotated ORCAS-I datasets and present the results in Table 10. To test the quality of Snorkel, we compare the distribution on ORCAS-I-gold both for manual annotations and the output from Snorkel weak labelling. We notice, that the only underrepresented category from our weak supervision is the navigational intent, which contains 1.6% less items than in manual labelling approach. These queries were mostly categorised as abstain by our weak supervision. The label distribution from Snorkel for all three datasets is comparable, so both gold and 2M samples chosen from the full ORCAS dataset are representative.

Table 10: Label distribution for all three annotated ORCAS-I datasets.

Label distribution	ORCAS-I-gold Manual	ORCAS-I-gold Snorkel	ORCAS-I-2M Snorkel	ORCAS-I-18M Snorkel
Navigational	17.10%	15.50%	14.48%	14.51%
Transactional	4.30%	4.50%	4.17%	4.16%
Informational	78.60%	80.00%	81.35%	81.33%
- Instrumental	5.90%	5.30%	5.82%	5.81%
- Factual	36.30%	35.40%	35.35%	35.35%
- Abstain	36.40%	39.30%	40.18%	40.17%

8 CONCLUSION

In this paper we revise the taxonomy of user intent, using the widely used classification of queries into navigational, transactional and informational as a starting point. We identify three different subclasses for the informational queries: instrumental, factual and abstain making the resulting classification of user queries more fine-grained.

Moreover, we introduce ORCAS-I, a new user intent classification dataset which extends the popular ORCAS dataset. The dataset is annotated using weak supervision with Snorkel. This approach

enables obtaining labels for all 18M query-URL pairs. It can be a suitable resource for altering retrieval system results to match the type of information request from the user. For example, for transactional queries search engines can put heavier weight on results with commercial content or sponsored links. By contrast, providing commercial results for factual queries should be avoided. The general domain and the size of the dataset, together with the taxonomy, allows multiple researchers to successfully train machine learning models to predict user intent to filter retrieval results.

Besides the annotated dataset, we also release our labelling functions that can be highly useful for future application. This also makes it easy to improve the labelling functions using feedback from other researchers and release an updated version of labelled datasets.

In addition to the the weakly supervised dataset, we also publish a manually annotated subset that can be used for benchmarking the quality of machine learning models. We test the accuracy of our weakly supervised annotations and compare them to five benchmark models showing that none of them is able to significantly outperform Snorkel's output.

One limitation of our study is that we fine-tuned the labelling functions based on the ORCAS dataset characteristics, such as the most commonly visited domains. URLs from the United States are over-represented in the ORCAS dataset. Users searching with the same query in another country, would be re-directed to a different website based on their location (especially for queries regarding medical and legal advice). Because there were no other click log datasets to our availability, we were not able to generalise the labelling functions to location-dependent URLs.

Future work will focus on improving the labelling functions to reach better generalisation on datasets with other location-aware features, and also extending the taxonomy to cover the exploratory intent.

We release all three annotated versions of the ORCAS-I dataset [26]. Moreover, for reproducibility and transparency, we make our data labelling and classification scripts publicly available on GitHub[6].

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval – DoSSIER (H2020-EU.1.3.1., ID: 860721).

REFERENCES

[1] Eytan Adar. 2007. User 4xxxxx9: Anonymizing query logs. *Proceedings of Query Log Analysis Workshop, International Conference on World Wide Web.*

[2] Azin Ashkan, Charles Clarke, Eugene Agichtein, and Qi Guo. 2009. Classifying and Characterizing Query Intent. 578–586. https://doi.org/10.1007/978-3-642-00958-7_53

[3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 876–885. https://doi.org/10.18653/v1/D16-1084

[4] Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Weak Supervision for Learning Discourse Structure. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2296–2305. https://doi.org/10.18653/v1/D19-1234

[5] Ricardo Baeza-Yates, Liliana Calderon-Benavides, and Cristina González-Caro. 2006. The Intention Behind Web Queries. *Lecture Notes in Computer Science* 4209, 98–109. https://doi.org/10.1007/11880561_9

[6] Michael Barbaro and Tom Zeller. 2006. A Face is exposed for AOL searcher no. 4417749. *New York Times* (01 2006).

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* 5 (7 2017), 135–146. http://arxiv.org/abs/1607.04606

[8] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36 (2002).

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[10] Katriina Byström and Preben Hansen. 2005. Conceptual framework for task in information studies. *JASIST* 56 (08 2005), 1050–1061. https://doi.org/10.1002/asi.20197

[11] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. Overview of the TREC 2014 Session Track. In *TREC*.

[12] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search. *arXiv preprint arXiv:2006.05324* (2020).

[13] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/3077136.3080832

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1 (10 2018), 4171–4186. https://arxiv.org/abs/1810.04805v2

[15] Alejandro Figueroa. 2015. Exploring effective features for recognizing the user intent behind web queries. *Computers in Industry* 68 (02 2015), 162–169. https://doi.org/10.1016/j.compind.2015.01.005

[16] Jason A Fries, Paroma Varma, Vincent S Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, et al. 2019. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature communications* 10, 1 (2019), 1–10.

[17] Sumeer Gul, Sabha Ali, and Aabid Hussain. 2020. Retrieval performance of Google, Yahoo and Bing for navigational queries in the field of "life science and biomedicine". *Data Technologies and Applications* (04 2020).

[18] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management* 44 (2008).

[19] Jim Jansen, Amanda Spink, and Sherry Koshman. 2007. Web Searcher Interaction With the Dogpile.com Metasearch Engine. *JASIST* 58 (03 2007), 744–755. https://doi.org/10.1002/asi.20555

[20] Jiepu Jiang, Daqing He, and James Allan. 2014. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time. *SIGIR '14: Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval* (07 2014). https://doi.org/10.1145/2600428.2609633

[21] In-ho Kang and Gilchang Kim. 2004. Query Type Classification for Web Document Retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* https://doi.org/10.1145/860435.860449

[22] Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. 2019. Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 4611–4621. https://doi.org/10.18653/v1/D19-1468

[23] Ashish Kathuria, Bernard J. Jansen, Carolyn Hafernik, and Amanda Spink. 2010. Classifying the user intent of web queries using k-means clustering. *Internet Research* 20 (2010).

[24] Melanie Kellar, Carolyn Watters, and Michael Author. 2007. A Field Study Characterizing Web-Based Information Seeking Tasks. *JASIST* 58 (05 2007), 999–1018. https://doi.org/10.1002/asi.20590

[25] Jeonghyun Kim. 2006. *Task as a predictable indicator of information seeking behavior on the Web.* Ph.D. Dissertation. Rutgers University.

[6] https://github.com/ProjectDoSSIER/intents_labelling

[26] Wojciech Kusa, Daria Alexander, and Arjen P. de Vries. 2022. ORCAS-I. https://doi.org/10.48436/pp7xz-n9a06

[27] Stefan Larson, Anish Mahendran, Joseph Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan Kummerfeld, Kevin Leach, Michael Laurenzano, Lingjia Tang, and Jason Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *EMNLP-IJCNLP 2019*. 1311–1316. https://doi.org/10.18653/v1/D19-1131

[28] Uichin Lee, Zhenyu Liu, and Junghoo Cho. 2005. Automatic identification of user goals in Web search. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*. 391–400. https://doi.org/10.1145/1060745.1060804

[29] Dirk Lewandowski. 2011. The retrieval effectiveness of search engines on navigational queries. *Aslib Proceedings* 63 (07 2011). https://doi.org/10.1108/00012531111148949

[30] Dirk Lewandowski, Jessica Drechsler, and Sonja Mach. 2012. Deriving query intents from Web search engine queries. *Journal of the American Society for Information Science and Technology* 63 (09 2012). https://doi.org/10.1002/asi.22706

[31] Yuelin Li. 2010. An exploration of the relationships between work task and interactive information search behavior. *JASIST* 61 (09 2010), 1771–1789. https://doi.org/10.1002/asi.21359

[32] Yumao Lu, Fuchun Peng, Xin Li, and Nawaaz Ahmed. 2006. Coupling Feature Selection and Machine Learning Methods for Navigational Query Identification. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. 682–689. https://doi.org/10.1145/1183614.1183711

[33] Gary Marchionini. 2006. Marchionini, G.: Exploratory search: from finding to understanding. Comm. ACM 49(4), 41-46. *Commun. ACM* 49 (04 2006), 41–46. https://doi.org/10.1145/1121949.1121979

[34] Alaa Mohasseb, Mohamed Bader-El-Den, and Mihaela Cocea. 2019. A customised grammar framework for query classification. *Expert Systems with Applications* 135 (2019), 164–180.

[35] Subhabrata Mukherjee, Ahmed Hassan Awadallah, and Jianfeng Gao. 2021. XtremeDistilTransformers: Task Transfer for Task-agnostic Distillation. arXiv:2106.04563 [cs.CL]

[36] 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. (11 2016).

[37] Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A Picture of Search. In *Proceedings of the 1st International Conference on Scalable Information Systems (Hong Kong) (InfoScale '06)*.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[39] Gustavo Penha, Alexandru Balan, and Claudia Hauff. 2019. Introducing MANtIS: a novel Multi-Domain Information Seeking Dialogues Dataset. (12 2019).

[40] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (04 2020), 8689–8696. https://doi.org/10.1609/aaai.v34i05.6394

[41] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proc. VLDB Endow.* 11, 3 (nov 2017), 269–282. https://doi.org/10.14778/3157794.3157797

[42] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/6709e8d64a5f47269ed5cea9f625f7ab-Paper.pdf

[43] Daniel Rose and Danny Levinson. 2004. Understanding User Goals in Web Search. *Thirteenth International World Wide Web Conference Proceedings, WWW2004* (04 2004). https://doi.org/10.1145/988672.988675

[44] Daniel Russell, Diane Tang, Melanie Kellar, and Robin Jeffries. 2009. Task Behaviors During Web Search: The Difficulty of Assigning Labels. In *2009 42nd Hawaii International Conference on System Sciences*. 1–5. https://doi.org/10.1109/HICSS.2009.417

[45] Abigail Sellen, Rachel Murphy, and Kate Shaw. 2002. How Knowledge Workers Use the Web. *CHI '02: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 4, 227–234. https://doi.org/10.1145/503376.503418

[46] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf

[47] Ryen White, Bill Kules, Steven Drucker, and m.c Schraefel. 2006. Supporting Exploratory Search, Introduction, Special Issue, Communications of the ACM. *Commun. ACM* 49 (04 2006).

[48] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[49] Guoqing Zheng, Giannis Karamanolakis, Kai Shu, and Ahmed Hassan Awadallah. 2021. WALNUT: A Benchmark on Weakly Supervised Learning for Natural Language Understanding. *arXiv preprint arXiv:2108.12603* (2021).