VoiSeR: A New Benchmark for Voice-Based Search Refinement

Simone Filice, Giuseppe Castellucci, Marcus Collins, Eugene Agichtein, Oleg Rokhlenko
Amazon
Seattle, USA
{filicesf,giusecas,collmr,eugeneag,olegro}@amazon.com

Abstract

Voice assistants, e.g., Alexa or Google Assistant, have dramatically improved in recent years. Supporting voice-based search, exploration, and refinement are fundamental tasks for voice assistants, and remain an open challenge. For example, when using voice to search an online shopping site, a user often needs to refine their search by some aspect or facet. This common user intent is usually available through a "filter-by" interface on online shopping websites, but is challenging to support naturally via voice, as the intent of refinements must be interpreted in the context of the original search, the initial results, and the available product catalogue facets. To our knowledge, no benchmark dataset exists for training or validating such contextual search understanding models. To bridge this gap, we introduce the first large-scale dataset of voice-based search refinements, VoiSeR, consisting of about 10,000 search refinement utterances, collected using a novel crowdsourcing task. These utterances are intended to refine a previous search, with respect to a search facet or attribute (e.g., brand, color, review rating, etc.), and are manually annotated with the specific intent. This paper reports qualitative and empirical insights into the most common and challenging types of refinements that a voice-based conversational search system must support. As we show, VoiSeR can support research in conversational query understanding, contextual user intent prediction, and other conversational search topics to facilitate the development of conversational search systems.

1 Introduction

Modern voice assistants, such as Amazon Alexa or Apple Siri, make use of Natural Language Understanding (NLU) techniques to perform several tasks. Some of the most popular functions offered by these systems are based on voice-search: millions use voice assistants to access information or search for music, products or local restaurants and stores. However, search experience with a voice assistant remains limited. The current generation of these systems mostly supports single-turn interactions, and does not naturally support more complex search needs, which often require *refinements* to narrow, broaden or change the initial search. Supporting refinement is a fundamental aspect of search systems, and it is done in a variety of ways in Web-based user interfaces, e.g., through query suggestion or explicit facets navigation or filtering. For example, in an e-Commerce search, a user may want to refine their search with respect to some facet or attribute (e.g., brand or price); this critical functionality is supported on most e-Commerce websites. However, this kind of interaction is challenging to support via voice-based dialogue interfaces, as interpreting such refinements requires modeling the original search intent, the initial results, and the available result facets.

To the best of our knowledge, no large scale dataset exists for training and validating NLU models for multi-turn voice-based search. To bridge this gap, we present a new Voice-based Search Refinement dataset, VoiSeR[1] to enable research into contextual understanding of multi-turn voice-based search. The VoiSeR dataset contains 9,810 utterances of voice refinements in the e-Commerce domain, paired with contextual information. Specifically, our dataset includes refinement utterances intended to filter results from a previous search with respect to some facet or attribute (e.g., brand, color, review rating, etc.).

The dataset was collected through crowdsourcing via Amazon Mechanical Turk, between February and June 2020 in the US and India. We designed the task to minimize any bias towards partic-

[1] The dataset is available for download at https://registry.opendata.aws/amazon-conversational-product-search/.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2321–2329
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

ular expressions or terminology which may not be natural to users. To achieve this goal, we provided clear and concise instructions; we intentionally did not provide examples to avoid biasing participants towards using specific linguistic expressions (see Figure 1 for an example). As a result, the dataset provides a diverse and natural representation of how users express the intended refinements during product search.

We annotated the dataset to highlight some important aspects characterizing a voice refinement of product search. In particular, we annotated (*i*) the products and attributes mentioned in each utterance, if present; (*ii*) the specific refinement intent of each utterance (e.g., refinement by exact attribute value).

In addition to the new VoiSeR benchmark dataset, our contributions include (*i*) an analysis of the data, where we highlight some linguistic aspects characterizing how people express the refinement intent, and (*ii*) an empirical investigation to demonstrates that VoiSeR can be successfully used to bootstrap NLU models for handling voice-based search refinements. Furthermore, we show that contextual information is beneficial for such NLU tasks.

Next, §2 provides details about the data collection and annotation. §3 provides a detailed analysis of the dataset, while §4 reports the empirical investigation. In §5, we discuss the related works. Finally, §6 discusses our conclusions.

2 VoiSeR Data Collection

In order to collect a large number of voice search refinements from multiple participants, we designed a crowdsourcing task on Amazon Mechanical Turk[2]. In §2.1, we provide details of the crowdsourcing experiment to collect refinement utterances. These utterances were annotated for intent and relevance, as described in §2.2.

2.1 Crowdsourcing the Voice Refinement Data

The design of the task was intended to make it both easy for the participants (i.e., Amazon Mechanical Turk Worker) and as realistic as possible, to provide valid linguistic expressions of voice refinements. Thus, we tried to reproduce a real "customer journey" of product searches and refinements. With this idea in mind, we designed the Mechanical Turk task depicted in Figure 1. The crowdsourcing interface shows to the Worker:

- An initial set of products, i.e., up to five products in the top part of the image.
- A target set of products, i.e., up to five products in the central part of the image.
- A visual intent indicator, i.e., an image describing the attribute type the worker should focus on when expressing the refinement. In Figure 1, this is the 5-star symbol on the central-left part; it represents the *review-rating* attribute.

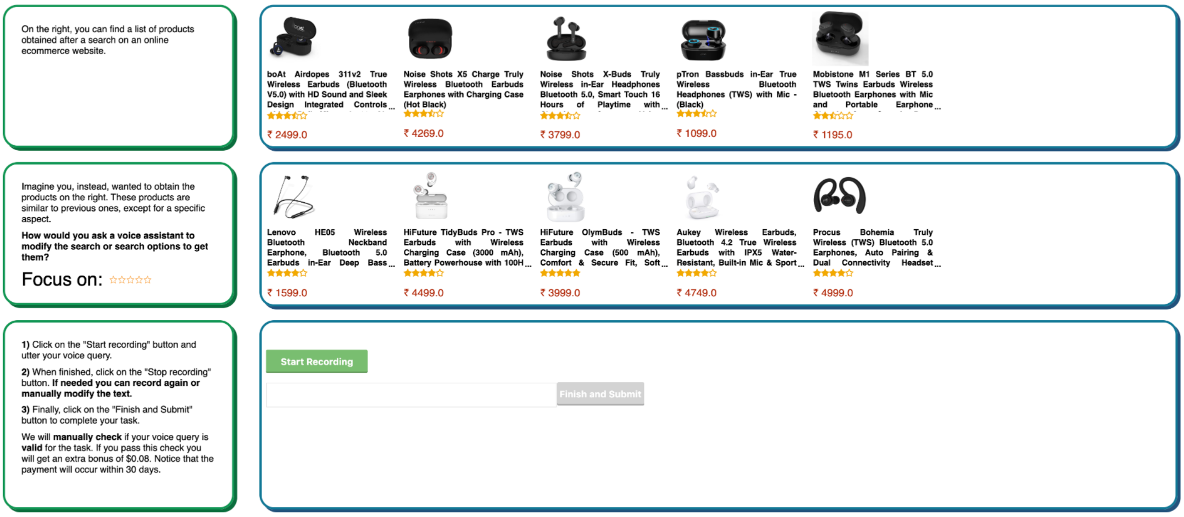
The participant is asked to imagine they are searching for products and that her search led to the initial set of results. We ask the participant to record a voice utterance, modifying the search to achieve the target product set, cued by the provided visual intent indicator. In the example in Figure 1, the participant should attempt to refine the search by *review rating* as indicated in the intent indicator: compared to the products in the initial set, the products in the target set have all 4 stars or more, therefore we expect the Worker to say something like "*Show me only products having 4 stars and above*" or "*Earphones with at least four stars*".

An Automatic Speech Recognition (ASR) system (we adopted Amazon AWS Transcribe[3]) processes the utterance in real time and its transcription is shown to the Worker. Since ASR errors can occur, before completing the task, we ask the Worker to check the ASR transcription. In case of errors, the Worker can record a new sentence or manually correct the ASR transcript. We record the original ASR transcript and any manual correction, if one is made.

To automatically generate the many examples to annotate, we used the Amazon.in product search engine: starting from a random product search, we collected the initially retrieved products, as well as those returned after the application of a filter. The type of the activated filter dictates the visual intent indicator shown to the Worker, while the products shown in the initial and target sets are a subset of those retrieved by Amazon.in before and after the filter application, respectively. To emphasize the difference between the initial set and the target set, we select the products so that (*i*) no product appears in both sets and (*ii*) the products in the

[2] https://www.mturk.com/

[3] https://aws.amazon.com/transcribe/



On the right, you can find a list of products obtained after a search on an online ecommerce website.

boAt Airdopes 311v2 True Wireless Earbuds (Bluetooth V5.0) with HD Sound and Sleek Design Integrated Controls ...
★★★⯪☆
₹ 2499.0

Noise Shots X5 Charge Truly Wireless Bluetooth Earbuds Earphones with Charging Case (Hot Black)
★★★⯪☆
₹ 4269.0

Noise Shots X-Buds Truly Wireless in-Ear Headphones Bluetooth 5.0, Smart Touch 16 Hours of Playtime with ...
★★★⯪☆
₹ 3799.0

pTron Bassbuds in-Ear True Wireless Bluetooth Headphones (TWS) with Mic - (Black)
★★★⯪☆
₹ 1099.0

Mobistone M1 Series BT 5.0 TWS Twins Earbuds Wireless Bluetooth Earphones with Mic and Portable Earphone ...
★★⯪☆☆
₹ 1195.0

Imagine you, instead, wanted to obtain the products on the right. These products are similar to previous ones, except for a specific aspect.

How would you ask a voice assistant to modify the search or search options to get them?

Focus on: ☆☆☆☆☆

Lenovo HE05 Wireless Bluetooth Neckband Earphone, Bluetooth 5.0 Earbuds in-Ear Deep Bass ...
★★★★☆
₹ 1599.0

HiFuture TidyBuds Pro - TWS Earbuds with Wireless Charging Case (3000 mAh), Battery Powerhouse with 100H ...
★★★★☆
₹ 4499.0

HiFuture OlymBuds - TWS Earbuds with Wireless Charging Case (500 mAh), Comfort & Secure Fit, Soft ...
★★★★★
₹ 3999.0

Aukey Wireless Earbuds, Bluetooth 4.2 True Wireless Earbuds with IPX5 Water-Resistant, Built-in Mic & Sport ...
★★★★☆
₹ 4749.0

Procus Bohemia Truly Wireless (TWS) Bluetooth 5.0 Earphones, Auto Pairing & Dual Connectivity Headset ...
★★★★☆
₹ 4999.0

1) Click on the "Start recording" button and utter your voice query.

2) When finished, click on the "Stop recording" button. **If needed you can record again or manually modify the text.**

3) Finally, click on the "Finish and Submit" button to complete your task.

We will **manually check** if your voice query is **valid** for the task. If you pass this check you will get an extra bonus of $0.08. Notice that the payment will occur within 30 days.

Figure 1: MTurk interface for voice refinements collection. The workers see source and target sets of products, differing on a specific attribute (e.g., review rating), and record a voice refinement by the suggested target attribute.

initial set do not satisfy the activated filter. For instance, the task shown in Figure 1 replicates a search session where the *4 stars & up* filter on the review rating was activated; therefore, for the initial set we selected all products having less than 4 stars, vs. more than 4 stars in the target sets.

We kept the task instructions as simple as possible in order to not introduce linguistic biases: the complete instructions to the Workers are those shown at the left Figure 1. We intentionally did not provide any example on how to complete a task. In relatively few cases this created some misunderstanding with the Workers that failed to provide valid voice refinements. On the other hand, our choice prevented any possible bias towards using some specific linguistic expressions. To further minimize terminology bias, we used visual intent indicators to suggest the attribute type to refine on. For instance, we used the image of a color palette to represent the *color* attribute, and an image of banknotes to suggest *price*.

In a preliminary experiment, we did not show the intent indicator, but in many cases the difference between the initial set and the target set was not obvious, so that the Workers ended up focusing on irrelevant details. As a consequence, the utterances that were collected in that setting were over-specific, and often the Workers simply read parts of the target product titles. Based on this experiment, the full dataset was collected using the intent indicator condition described above.

In e-commerce websites selling a wide range of products, there are typically many possible attribute types customers can filter on. We decided to collect data about some of the most popular and generally applicable ones, namely *brand, color, discount, material, price* and *review rating*.

2.2 Data Annotation

After collecting the data, we asked domain experts to annotate them with respect to three different tasks.

Voice Refinement Validity: Since crowdsourcing data can be noisy, we first design a preliminary annotation task to validate each single utterance collected on Mechanical Turk. In particular, we showed to the annotators the Mechanical Turk task associated to each sentence and asked them to state whether the utterance correctly refines the product search on the target attribute. We also asked the annotators to report whether ASR errors (or typos in case the Worker manually corrected the ASR transcription) occur in the utterance.

Attribute and Product Extraction: We asked the annotators to mark product and attribute mentions in the utterance, for example in the utterance *"Show me only red t-shirts," red* is the attribute and *t-shirt* is the product. Given the Mechanical Turk task design discussed in §2.1, the collected utterances are supposed to refine on a single attribute. However, we noticed that some utterances contain multiple attribute mentions. For instance, *Nike* and *red* in the sentence *"Show me only red Nike t-shirts"*. The annotators are required to extract all individual attribute mentions within the utterance and not only the attribute mention referring to the target attribute type (the one shown as intent indicator to the Worker).

Refinement Intent Classification: This task consists of indicating how to change the search query based on the attribute mentioned in the utterance. We asked the annotators to indicate whether the provided refinement belongs to one of the following types:

- EXACT: the customer asks to select products having a specific value for the attribute, e.g., *"show me only purple"*.
- EXCLUDE: the customer asks to exclude products having a specific value for the attribute, e.g., *"exclude the purple ones"*.
- RANGE: the customer asks to select products having attribute values in a closed interval, e.g., *"Price between 200 and 300"*.
- GREATER: the customer asks to select products with attribute value higher than a given value, e.g., *"Show 4 stars and up"*, or *"Exclude the products with less than four stars"*.
- LOWER: the customer asks to select products with attribute value lower than a given value, e.g., *"Price less than 100"*, or *"Exclude the ones more expensive than 100"*.
- OTHER: utterance not falling in the above categories, e.g., *"Show me a different color"*, or *'Select top ratings"*.

Each example in our data is annotated by a single domain expert, since we observed a very high annotation quality in a preliminary annotation phase where multiple annotators annotated the same instances. We registered an almost perfect agreement in all tasks: Cohen's Kappa 0.914 for the Voice Refinement Validity task, Cohen's Kappa 0.960 for the Attribute and Product Extraction task, and Cohen's Kappa 0.859 for the Refinement Intent Classification task.

3 Voice Refinements Dataset

In this section, we provide the analysis of the data collected through the crowdsourcing experiment. First, we describe how we conducted the Mechanical Turk experiment and the statistics of the collected data in §3.1; then, we discuss some of the linguistic properties emerging in the context of voice refinements in §3.2.

3.1 Crowdsourcing the Voice-Based Search Refinement Data

The data was crowdsourced using the Amazon Mechanical Turk platform, from workers in the U.S. and India, between February and June 2020. Both Indian and U.S. Workers were asked to provide English voice refinements with respect to the tasks shown.

Region	**Total**	**Valid (%)**
US	2,716	2,475 (91.12%)
IN	10,776	7,335 (68.06%)
Total	13,492	9,810 (72.71%)

Table 1: Data statistics per region.

As reported in Table 1, we collected 13,492 utterances, and 9,810 (i.e., ~72%) were considered valid refinements in the subsequent annotation phase, while the rest are not refinements or contain ASR errors. In total 385 workers participated in the data collection, each one producing 35 utterances on average[4]. This results in a great variability in the collected data, as each worker can provide different linguistic expressions for a voice refinements.

Table 2 reports the distributions of the valid utterances with respect to the attribute types and the refinement types. The utterances are mostly evenly distributed with respect to the 7 attributes. *Brand* and *material* are the most and least represented attributes. The majority of utterances are of type EXACT, i.e., they ask to filter the product search on a specific attribute value. Unsurprisingly, the EXCLUDE type is extremely rare: our Mechanical Turk task did not encourage this refinement type.

As expected, the majority of brand-related utterances are of type EXACT. Also, the number of utterances of the price and rating related attributes are mainly LOWER and GREATER, respectively. This is intuitive considering that usually users look for less expensive and better rated products (for example, "*Show me items that are rated four stars or better*"). Notice that price (and also discount) utterances include a good number of RANGE, EXACT, as for example, "*Womens bags in the range of 1000 to 3300*".

Most of the OTHER bucket are (*i*) utterances whose intent is to sort products with respect to an attribute type like price or review rating, e.g., "*Sort by most rated laptops*", (*ii*) utterances where no specific attribute value is mentioned, e.g., "*Show*

[4] Each worker was allowed to do at most 100 tasks.

	EXACT	**EXCLUDE**	**GREATER**	**LOWER**	**RANGE**	**OTHER**	**Total**
brand	2,951	1	1	0	1	26	2,980
color	1,321	1	0	0	0	201	1,523
discount	480	0	234	80	37	478	1,309
material	719	2	0	0	0	19	750
price	119	0	176	799	134	0	1,463
rating	401	0	852	7	18	517	1,795
Total	5,991	4	1,263	886	1,477	189	9,810

Table 2: Distribution of the valid utterances with respect to the attribute types and the refinement types.

items with high ratings”. Other common cases are utterances with comparatives e.g., “*Show me less expensive earphones*” or superlatives e.g., “*Show me the highest rated silk sets*”.

We noticed that, despite the fact that our Mechanical Turk tasks focused on a single attribute, in about 27% of the cases (2,668 utterances) Workers provided utterances containing multiple attributes. For example, the utterance “*Price should be less than 700 with discount*” contains both the target attribute (discount) and a specification of the price.

Finally, ∼80% of the utterances have a product mention, e.g., “*Show me a Vega brand hair curler*”. The rest don’t mention products, e.g., “*Which ones are discounted*”, or “*Higher price*”.

3.2 Descriptive Analysis

To better understand the collected data, and to identify differences between regions, categories, *etc*..., we computed some basic features of the refinement utterances: word counts, entropy per word for a bi-gram language model, use of adjectives and modifiers, dependency tree depth, and whether the utterance could be parsed as a complete English sentence having subject, verb, and object. We find moderate but statistically significant differences between regional populations which may have implications for the applicability of this data set to other regions. Table 3 shows the statistically significant differences. We found that utterances from IN had a much longer tail on many of these statistics. While the mean word count is significantly smaller in IN than the US, the maximum word count in IN is nearly three times that in the US.

Adjectives and modifiers are used much more in IN than US, which appears to reflect workers speaking not only the refinement, but the original product search query as well. For example “*I’m looking for a blue straight fit trouser pant*” includes much more than the target color refinement, “*blue*”. US workers had an increased tendency to speak in complete sentences, e.g., “*Please display more perfumes from Calvin Klein*” instead of simply “*Calvin Klein*”, both of which are utterances found in the dataset. Perhaps the most curious distinction is US workers’ increased tendency to use first-person pronouns compared to other workers. These are typically phrased as “*I only want to see...*”, “*Show me...*”, “*I want the* {*attribute*} *to be* {*value*}”, “*I’m looking for ...*” etc.

Region	**US**	**IN**
word count	7.2	5.7
adjective/modifier count	0.82	1.0
mean tree depth	1.8	1.5
per-word entropy	0.68	0.94
% using first-person pronouns	16	2.7
% complete sentences	23	17

Table 3: Differences between utterances from different regions. All differences are statistically significant ($p < 0.01$, with a Tukey multiple-comparison correction).

Unsurprisingly, the per-word entropy is quite a bit higher for brand refinements than any other type but rating, while the word count is smaller for brand refinements than most others. In brand refinements the usage of first person pronouns is higher than any other type. Ratings refinements are the longest overall, with the deepest dependency trees.

We note that the distribution of refined attributes differs across regions. With one notable exception, we do not see any statistically significant, consequential, and consistent interactions between region and refined attribute when used to predict the features shown in Table 3. That exception is that review rating refinements are even longer and have even more per-word entropy in the US than IN than would be predicted otherwise. However, these findings only exaggerate trends that are already present in the data overall. This largely serves to suggest that review ratings are some of the most complex in the dataset.

4 Empirical Investigation

In this section we present a set of experiments on intent detection, specifically the recognition of attribute and product categories in a search refinement utterance. We aim to show that the VoiSeR dataset enables building models for intent recognition in voice search refinement. Moreover, we show the contribution of contextual information, e.g., the previous utterance, is beneficial, highlighting the need for large scale datasets like VoiSeR for developing contextual intent recognition models.

Set	# Utterances	# Workers
Train	8,432	337
Validation	653	28
Test	725	20

Table 4: Data distribution and number of workers in the training, validation and test splits.

Experimental setup. In order to train and test the model, we split the data into `train`, `validation` and `test` portions, as reported in table 4. The splitting has been done at Worker level, i.e., utterances from each Worker appears only in a single portion.

Model. We model the recognition of attribute and product mentions in a refinement utterance as a sequence tagging problem in an IOB2[5] tagging schema. For example, given the utterance "*Show me a Vega brand hair curler*", its correct tagging is "`O O O B-Attribute I-Attribute B-Product I-Product`". Notice that we do not tag the attribute type, but only a general category `Attribute`. This is to enable the model to generalize over the linguistic expressions that are similar among different attribute types. The model we implemented is a BERT-based (Devlin et al., 2019) sequence tagger. BERT is used as the encoder to obtain the contextualized embeddings of each token w_i of a sentence, i.e., $h_i = BERT(w_i)$. After applying dropout on h_i, a linear classifier is used to obtain the c_i distribution over the IOB categories for each token w_i, i.e., $c_i = softmax(Wh_i + b)$, where softmax refers to the function to transform the scores in output probabilities and W and b are the weights and bias of the classifier, respectively. The model is trained by optimizing the cross entropy between the predicted categories and the true one-hot distributions for each token. We run experiments on two settings: (*i*) w/o context, i.e., the input to the model is only the refinement utterance; (*ii*) w/ context, i.e., the model receives as input both the original search query and the refinement utterance. The two utterances are separated by the `[SEP]` token, as usual in BERT (Devlin et al., 2019). The latter experiment will provide insights about the utility of contextual information to the model.

Experimental Setup. We used the `bert-base-uncased` model from the Huggingface repository (Wolf et al., 2019) as the encoder. The model is trained for 15 epochs with Early Stopping (patience=3) by tuning the following hyperparameters: dropout (0.0, 0.1) applied on the h_i representations; batch size (64, 128). The learning rate is set to 5e-5, with the Adam Optimizer with a warm-up of 100 steps.

Experimental Results. Table 5 reports the model performance (with and without context) at entire span level and at token level. The former measures the capability of the model in recognizing an entire attribute or product, i.e., the attribute or product are considered correctly predicted only if all their tokens are recognized by the model. The latter measures the capability of the model in partially recognizing an attribute or product in a voice refinement. The model performance is already promising when context is not leveraged, achieving 83.11 and 84.36 F1 for attribute and product, respectively. Contextual information provides a significant improvement, allowing the model to reach 84.94 and 86.86 F1 for attribute and product, respectively. This demonstrates that contextual information helps NLU models to better understand customer requests.

We performed an in-depth analysis to find out how the model performs on different attribute types. Tables 6 and 7 report results for each attribute type[6] in the w/o and w/ settings, respectively.

In both settings, the model achieves the best results on *discount*, while results are a bit worse on *material*, *review-rating* and *brand*. This is a consequence of the lower linguistic variability associated with the *discount* attribute. On the other hand, refinements on *review-rating* are on average the

[5] IOB is a short for Inside, Outside, Beginning. In the IOB2 schema, the B- prefix before a tag indicates the beginning of an entity. For subsequent words of an entity the I- prefix is used. Words not belonging to any entity are tagged with O.

[6] Again, we do not tag the attribute type; we report the measures w.r.t. the target attribute used in the Mechanical Turk task.

Model		P	R	F1
w/o ctx	**Attr**	81.28 (95.47)	85.01 (95.27)	83.11 (95.37)
	Prod	83.08 (**94.22**)	85.67 (94.39)	84.36 (94.30)
w/ ctx	**Attr**	**83.71 (96.05)**	**86.21 (95.38)**	**84.94**† **(95.71)**
	Prod	**86.20** (94.16)	**87.54 (94.83)**	**86.86**† **(94.49)**

Table 5: Experimental results on the attribute and product extraction task. The sequence tagging models is tested when using or not the context. Reported metrics are Precision (P), Recall (R) and F1 at entire span level (in parenthesis the same metrics at token level). The symbol † marks a statistical significant difference with a paired t-test ($\alpha = 0.05$, for both Attribute and Product categories the p-value is ~ 0.009).

Attribute		Precision	Recall	F1
brand	**Attr**	79.45	83.40	81.38
	Prod	80.37	84.52	82.39
color	**Attr**	83.17	83.57	83.37
	Prod	81.62	87.40	84.41
discount	**Attr**	84.21	89.51	86.78
	Prod	91.40	90.43	90.91
material	**Attr**	79.82	84.26	81.98
	Prod	92.75	88.89	90.78
price	**Attr**	88.77	90.22	89.49
	Prod	78.57	80.49	79.52
rating	**Attr**	69.17	77.97	73.31
	Prod	80.00	84.51	82.19

Table 6: Span-level Precision (P), Recall (R) and F1 of the model when computing the performances by dividing the test set on the target attribute used in the data collection (w/o context).

Attribute		Precision	Recall	F1
brand	**Attr**	78.82	83.40	81.05
	Prod	84.62	85.16	84.89
color map	**Attr**	84.29	85.51	84.89
	Prod	85.61	88.98	87.26
discount	**Attr**	85.33	89.51	87.37
	Prod	93.55	92.55	93.05
material	**Attr**	80.91	82.41	81.65
	Prod	88.89	88.89	88.89
price	**Attr**	89.78	90.76	90.27
	Prod	81.75	83.74	82.73
review rating	**Attr**	84.17	85.59	84.87
	Prod	86.30	88.73	87.50

Table 7: Span-level Precision (P), Recall (R) and F1 of the model when computing the performances by dividing the test set on the target attribute used in the data collection (w/ context).

longest and have the highest linguistic variability, as discussed in §3.2. Similarly, *brand* refinements contain a lot of rare words, i.e., the brands. Finally, a possible explanation for the lower performance on the *material* refinements is the smaller training size, as shown in Table 2. Less variability w.r.t. the attribute type of the refinement is observed on the product recognition. Again, the performance is generally higher when using the previous utterance as context. In this case, the model is able to make better predictions especially for *rating* (+12 F1 points).

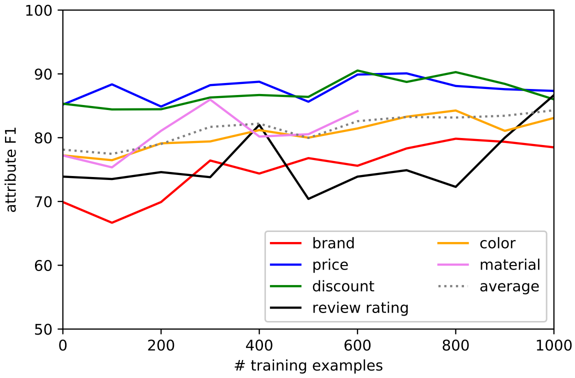


Figure 2: Learning curves for the attribute extraction task. The curve for attribute type a is obtained by training the model on all the utterances in the training set targeting a different attribute and on an increasing number of utterances targeting the attribute a.

Finally, we conducted a set of experiments to study the model capability to generalize to attribute types rarely, or never, observed in training. Figure 2 reports a learning curve for each attribute type we collected. The learning curve is computed as follows: for each reported attribute a and for each size s in the x-axis, we train a context-based model with s utterances of the attribute[7] a and with all the utterances for the other attributes. In figure 2 we report the F1 of the attribute recognition for each attribute and also the average F1 for all the attributes at each size s. The average F1 score on never seen attribute types (i.e., when using $s = 0$ examples for each attribute) is ~70: even with no training examples, the model obtains good results on all attributes. In terms of generalization capability, the best performance is observed on $price$ and $discount$, with learning curves that start from high scores and remain almost flat. This is not surprising, since these two attribute types are related, and the model can easily generalize from one to the

[7] The learning curve for *material* ends earlier because there are fewer utterances in the training set.

other. The *brand* attribute improves more when increasing the number of training examples. This may be due to the large lexical variability in this attribute: in e-commerce catalogues the number of brand names is typically very large.

The average learning curve becomes almost flat after 400 examples, and overall the model demonstrates a very good generalization capability on new attribute types. This suggests that the collected dataset has a reasonable size and that it represent a valuable resource to bootstrap NLU models for voice-based search.

5 Related Work

Human-computer information retrieval (HCIR) (Marchionini, 2006) combines the fields of human-computer interaction (HCI) and information retrieval (IR) to create systems that improve search by taking into account the human context, or through a multi-step search process that provides the opportunity for human feedback. Modern search engines implement several HCIR strategies including relevance feedback (Ruthven and Lalmas, 2003), automatic query completion (Cai and de Rijke, 2016) and faceted search (Tunkelang, 2009). In particular, most of the e-commerce websites provide faceted search by allowing users to filter on several product attributes. With our work we would like to study how a similar experience can be made available in a voice-based interaction.

Due to the increasing availability of smartphones and voice assistants like Amazon Alexa or Google Home, voice-based search is becoming ubiquitous (Guy, 2016). Voice-based Web search is significantly different from text-based search (Guy, 2016, 2018), notably for voice-based query reformulation or refinement (Hassan Awadallah et al., 2015). The voice-based interaction is especially challenging when no visual interface is available and both inputs and outputs are entirely provided by voice (Ingber et al., 2018). To improve user experience and engagement in voice-search some work has been recently done. Gamzu et al. (2020) propose a query rewriting approach for handling mispronounced, misexpressed, and misunderstood customer queries in voice shopping. Filice et al. (2020) describe the problem of reformulating answers from a community question answering system to make them suitable for a voice interaction.

There have been prior efforts in creating open multi-turn voice-based search datasets, but because of the lack of effective automated systems for these tasks, the datasets were collected in a lab using a Wizard-of-Oz approach, where a hidden human participant playing the part of the search engine, e.g., (Trippas et al., 2017, 2020; Vakulenko et al., 2019), and (Trippas and Thomas, 2019). The dataset we contribute complements prior efforts by providing a large-scale collection of voice-based query reformulation, collected in a realistic environment with a real search engine, using automated ASR, thus providing a critical resource for training robust, high-performance models for voice refinement.

Recently, several dialog-related datasets have been (Mehri et al., 2020; Crook et al., 2019) proposed. They target task-oriented conversations but do not specifically focus on search tasks.

6 Conclusions

In this paper, we discussed the problem of search refinement, a fundamental component for supporting multi-turn voice-based complex search tasks. We presented the challenges in the voice refinement problem, and introduced a large-scale, critically needed benchmark for training and evaluating models in this setting. Specifically, we introduced the first benchmark dataset, VoiSeR, specifically developed for analyzing and measuring the linguistic phenomena underlying the voice refinements in second-turn searches in the e-commerce domain. We emphasize that the target search facets and attributes are (by design) general, and thus the data and the resulting models can be used, with or without adaptation, for a wide range of conversational search refinement and intent prediction tasks.

We provided a detailed description of the data collection and annotation processes, and identified interesting statistical and linguistic phenomena in the dataset. We complement the data release with an extensive empirical investigation, which demonstrates that (*i*) our dataset is a valuable resource for training NLU models for voice-based search and (*ii*) using contextual information for recognizing product and attribute mentions is beneficial. Together, the new VoiSeR dataset and the analysis in this paper enable productive research for developing systems for voice-based complex search tasks.

References

Fei Cai and Maarten de Rijke. 2016. *A Survey of Query Auto Completion in Information Retrieval*. Now Publishers Inc., Hanover, MA, USA.

Paul A Crook, Shivani Poddar, Ankita De, Semir Shafi, David Whitney, Alborz Geramifard, and Rajen Subba. 2019. Simmc: Situated interactive multimodal conversational data collection and evaluation platform. *arXiv preprint arXiv:1911.02690*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Simone Filice, Nachshon Cohen, and David Carmel. 2020. Voice-based reformulation of community answers. In *Proceedings of The Web Conference 2020*, WWW '20, page 2885–2891, New York, NY, USA. Association for Computing Machinery.

Iftah Gamzu, Marina Haikin, and Nissim Halabi. 2020. Query rewriting for voice shopping null queries. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20.

Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 35–44, New York, NY, USA. Association for Computing Machinery.

Ido Guy. 2018. The characteristics of voice search: Comparing spoken with typed-in mobile web search queries. *ACM Transactions on Information Systems (TOIS)*, 36(3):1–28.

Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and predicting voice query reformulation. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 543–552.

Amir Ingber, Arnon Lazerson, Liane Lewin-Eytan, Alexander Libov, and Eliyahu Osherovich. 2018. The challenges of moving from web to voice in product search. In *Proceedings of the 1st International Workshop on Generalization in Information Retrieval*.

Gary Marchionini. 2006. Toward human-computer information retrieval. *Bulletin of the American Society for Information Science and Technology*, 32(5):20–22.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue.

Ian Ruthven and Mounia Lalmas. 2003. A survey on the use of relevance feedback for information access systems. *Knowledge Eng. Review*, 18(2):95–145.

Johanne Trippas and Paul Thomas. 2019. Data sets for spoken conversational search. In *Workshop on Barriers to Interactive IR Resources Re-use at the ACM SIGIR Conference on Human Information Interaction and Retrieval*.

Johanne R Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. A conversational search transcription protocol and analysis. In *Proceedings of SIGIR 1st International Workshop on Conversational Approaches to Information Retrieval (CAIR'17), CAIR*.

Johanne R Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a model for spoken conversational search. *Information Processing & Management*, 57(2):102162.

Daniel Tunkelang. 2009. *Faceted Search*. Morgan and Claypool Publishers.

Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. Qrfa: A data-driven model of information-seeking dialogues. In *European Conference on Information Retrieval*, pages 541–557. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.