A Dataset for Sentence Retrieval for Open-Ended Dialogues

Itay Harel*
TSG IT Advanced Systems Ltd.
Tel Aviv, Israel
itay.harel91@gmail.com

Hagai Taitelbaum
Google Research
Tel Aviv, Israel
hagait@google.com

Idan Szpektor
Google Research
Tel Aviv, Israel
szpektor@google.com

Oren Kurland
Technion — Israel institute of technology
Haifa, Israel
kurland@ie.technion.ac.il

ABSTRACT

We address the task of sentence retrieval for open-ended dialogues. The goal is to retrieve sentences from a document corpus that contain information useful for generating the next turn in a given dialogue. Prior work on dialogue-based retrieval focused on specific types of dialogues: either conversational QA or conversational search. To address a broader scope of this task where any type of dialogue can be used, we constructed a dataset that includes open-ended dialogues from Reddit, candidate sentences from Wikipedia for each dialogue and human annotations for the sentences. We report the performance of several retrieval baselines, including neural retrieval models, over the dataset. To adapt neural models to the types of dialogues in the dataset, we explored an approach to induce a large-scale weakly supervised training data from Reddit. Using this training set significantly improved the performance over training on the MS MARCO dataset.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Retrieval tasks and goals**;

KEYWORDS

dialogue retrieval; sentence retrieval

ACM Reference Format:
Itay Harel, Hagai Taitelbaum, Idan Szpektor, and Oren Kurland. 2022. A Dataset for Sentence Retrieval for Open-Ended Dialogues . In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477495.3531727

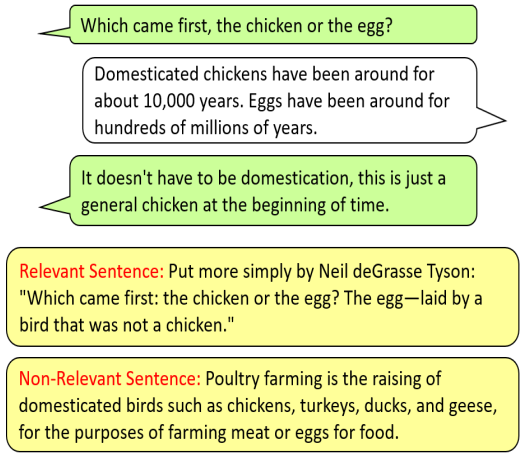


Figure 1: Test example of a dialogue created from Reddit with sentences judged as relevant and non-relevant by human annotators.

1 INTRODUCTION

There has been a rapid increase in the last few years in research of tasks related to dialogue (conversational) systems [8, 12, 14, 15, 35, 50]. Prominent examples include response generation [12, 16, 31, 55] or response selection [3, 17, 33, 36, 39, 44, 45, 53] with respect to the last turn in the dialogue, conversational question answering [13, 23, 55] and conversational retrieval (of passages) [8, 15, 34, 50].

In this paper we focus on open-ended dialogues: two parties converse in turns on any number of topics with no restrictions to the topic shifts and type of discussion on each topic. In addition, the dialogue is not grounded to a specific document, in contrast to the setting used in some previous work (e.g., [25]). The task we address is retrieving sentences from some document corpus that contain information useful for generating (either automatically or by humans) the next turn in the dialogue. We note that the dialogue turns can be questions, queries, arguments, statements, etc.

Existing dialogue/conversational datasets are not well suited for evaluation of the task we pursue; we discuss this point in length in Section 2. Hence, we developed a novel dataset reported here. The dataset includes 846 dialogues created from Reddit threads. For each dialogue, 50 sentences were retrieved from Wikipedia using an unsupervised initial retrieval method. These sentences were judged by crowd workers for relevance, that is, whether they contained information useful for generating the next turn in the

*Work done while at the Technion.

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

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531727

dialogue. Figure 1 depicts one such dialogue, with two sentences annotated by the raters, one as relevant and one as non-relevant. The dataset is available at https://github.com/SIGIR-2022/A-Dataset-for-Sentence-Retrieval-for-Open-Ended-Dialogues.git.

Neural-based retrieval methods require lots of training data — whether for learning from scratch or for fine tuning a pre-trained model. Hence, we used a weak supervision approach to induce pseudo relevance labels for a few sentences for each of ~73,000 additional dialogues. To this end, we fused rankings induced by several methods over an initial sentence list. These methods are either unsupervised or are based on distant supervision. For example, we used a BERT-based model [11] trained for query-based passage retrieval on the MS MARCO dataset [29].

We report retrieval performance over the dataset for several methods. The evaluation demonstrates the clear merits of using the pseudo relevance labels induced using the weak supervision to fine tune BERT-based retrieval methods.

Our contributions can be summarized as follows:

- A dataset for open domain dialogue systems with data labeled for the task of retrieving sentences that contain useful information for response generation.
- A procedure to induce a large scale weakly supervised training data using the metadata of Reddit dialogues.

2 RELATED WORK

There are two main lines of work related to ours. The first is on devising datasets for conversation systems. The second line is on using weak supervision for retrieval.

2.1 Conversational Datasets

Existing conversational datasets were built for two main tasks. The first is to compose the next turn in a dialogue, either via generative language models or by retrieving full responses from an index. Therefore, related datasets [1, 9, 12, 16, 21, 24, 26, 32, 42, 44, 47, 52, 54] serve to evaluate the offered responses compared to gold responses, not the retrieval of relevant information for composing such responses. The second task focuses on conversational passage retrieval and question answering (QA), where information needs are conveyed by a user to a search engine or a question answering system via a series of queries/questions that should be considered a single session. Prominent conversational QA datasets include CoQA [35], DoQA [2] and QuAC [3]. In these datasets, all evaluation sessions are grounded to a single passage or section that the participants are allowed to ask and answer about. In contrast, we address dialogue-based sentence retrieval from a document corpus.

The only conversational passage retrieval dataset we are familiar with is from TREC's CAsT tracks [7, 8]. However, CAsT's queries reflect explicit intents, while we are also interested in more open dialogues where the information needs can be in the form of implicit intents, as shown for example in Figure 1. In these datasets, the user conducts a query session on a specific single topic. The queries in the session may co-refer and reflect prior queries in the session. However, in most of these datasets, the returned search results are not viewed as part of the dialogue. Finally, in both conversational passage retrieval and conversational QA datasets, there is a user asking questions or queries that reflect explicit intents with information needs, as opposed to natural dialogues where intents may be only implicitly represented, e.g., in affirmative statements.

To sum, existing conversational datasets do not combine natural human-human conversations with relevance annotations for sentences retrieved from a large document corpus. We therefore constructed such a dataset and present it in Section 3.

2.2 Weak Supervision for Retrieval

A standard approach for using neural models in IR (and text-based tasks in general) , specifically dense retrieval models [22], is to first pre-train the neural model, either on a different target task but with a lot of training data, such as masked language modeling, or on a similar task as the end-goal, but on a different domain. Then, the pre-trained model is fine-tuned on training data for the target task or domain [11, 18, 29]. For example, Yilmaz et al. [46] fine-tuned a sentence retrieval model by first training a retrieval model on the MS MARCO retrieval task, and then fine-tuning it on a Microblog retrieval dataset.

Fine-tuning of retrieval models requires relevance labels for training examples in a target task. These are sometimes scarce or unavailable. One approach to circumvent this is to automatically generate labels and train a weakly supervised model on these annotations. Wu et al. [43] trained a response selection model for conversations using a weak signal induced from a matching probability offered by a seq2seq model trained on human conversations. Li et al. [20] used weak annotators based on search logs to train a Search Ads matching model. They automatically selected pseudo negative examples by optimizing the distance between a pseudo negative example from a retrieved pool and a given positive example. Dehghani et al. [10] used Okapi BM25 rankings to induce pseudo relevance labels so as to train a neural model ranker. Zamani and Croft [48] then provided a theoretical analysis for the merits of using such weak supervision for ad hoc retrieval. Weak supervision was also used for other related retrieval tasks; for example, expanding the last dialogue turn for passage retrieval [40] and for query performance prediction [49].

We follow the weak supervision paradigm in our model training, with a novel weak Reddit annotator for retrieval in a dialogue context.

3 A DATASET FOR OPEN DIALOGUE RETRIEVAL

As described in Section 2.1, we are not aware of a dataset that can be used to evaluate retrieval in an open-ended dialogue context. We therefore constructed such a dataset. To this end, we used dialogues from Reddit, due to its large variety of topics and styles. We then retrieved candidate Wikipedia sentences for these dialogues. Each candidate was judged for relevance by crowd-source human annotators; it was marked as relevant if it contained information useful for generating the next turn in the dialogue. We next detail the way we collected dialogues from Reddit and the manual annotation process and construction of the dataset.

3.1 The Reddit Collection

Reddit is a multi-party discussion platform, where users add turns in a conversation that can be replied to by several users. Each discussion is therefore a tree structure of turns, and a path from the initial turn (the tree's root) to a turn without followup replies (a leaf) is called a *thread*. We collected 263, 986, 046 conversation threads from Reddit after preprocessing the submission and comment files provided by the Pushift.io archives[1]. All threads are in English, none of them belongs to topics (called *subreddits* in Reddit) on offensive content, and they do not contain missing turns, i.e., turns that have been removed and their text is not displayed. We refer to the collection of these threads as the *Reddit collection*.

We enriched the context of each thread by prepending to the first turn: (a) the name of the subreddit the discussion was issued under, and (b) the title of the discussion (if provided by the user who initiated the thread). We split the threads by dates: test set candidates (discussed in Section 3.2) were limited to dates between Feb-2011 and Dec-2013, and training set candidates (discussed in Section 4.2) were limited to dates between Jan-2014 and Aug-2019.

Some turns in Reddit offer one or more links to Web pages that the author considers related to the written text. We only considered links to specific sections on Wikipedia pages that are found in our downloaded version of Wikipedia; we refer to these turns as *Wikipedia grounded turns*.

3.2 Test Set Construction

To construct a test set, we randomly sampled threads from the test set part of our Reddit collection which was described in Section 3.1. A single thread in Reddit represents a multilogue. Hence, to distill dialogue-like conversations from threads, we considered only threads in which the user who initiated the discussion, called here the *initiator*, is the author of each odd turn of the thread. All other users in the thread, who author the even turns, are considered the *responders*. These threads are thus interleaving turns between the initiator and the responders. Threads with less than 4 turns were discarded, since we want a meaningful dialogue context. We refer to the remaining threads as *Reddit dialogues*. In each candidate dialogue, we refer to the last responder turn as the *target turn*. A *test dialogue* was constructed by trimming the candidate dialogue to include turns up to, but excluding the *target turn*.

We used an initial sentence retrieval method (henceforth also referred to as an initial ranker), described in Section 3.3, to retrieve 50 candidate sentences from Wikipedia for each test dialogue. The retrieval method utilizes all turns in the test dialogue. We then recruited master[2] workers from Amazon Mechanical Turk to judge the relevance of each retrieved sentence. Our instructions, presented in Figure 6, were to mark a sentence as relevant to a dialogue if it contained information useful for generating the next turn in the dialogue that would be a natural continuation. We provided the annotators with the dialogue turns, including its topic and title, and the Wikipedia sentences including their page's title. The definition of sentence relevance — as conveyed to the annotators — does not indicate the correctness or soundness of a turn that might be generated based on information in the sentence.

[1]https://files.pushshift.io/reddit/

[2]https://www.mturk.com/worker/help#what_is_master_worker

Table 1: Test set statistics (average, median and standard deviation) of the number of relevant sentences and the rank of the first relevant sentence (the highest rank is 1) retrieved by the initial ranker per dialogue.

Dialogue type	Ungrounded			Wikipedia Grounded		
	Avg	Med	Std	Avg	Med	Std
# Relevant	4.43	3	4.45	5.68	4	5.09
Rank of 1st Relevant	11.3	6	12.5	9.395	4	12.14

Every Mechanical Turk worker had to annotate 10 dialogues and 5 retrieved sentences for each dialogue. We truncated the turns and sentences to the first 30 words for ease of use; the workers had access to the whole text by hovering the cursor over it. At most three workers judged a sentence for relevance. We started with a majority vote between two workers, and if no majority agreement was achieved, the third vote was then used. We used honeypots to identify and disqualify raters that seemed to not follow our guidelines. An example dialogue from our test set with one sentence judged as relevant and one as non-relevant is shown in Figure 1.

Since we show candidate test set dialogues to humans, we filtered out test dialogues whose content includes one or more words from a public list of dirty words[3]. We also expanded this list by concatenating phrases in the list to detect hashtags and titles consisting dirty words without spaces. In addition, we filtered out dialogues whose turns contain URL references (with the exception of the Wikipedia grounded target turns, which are not shown to the raters), and those with turns shorter than 5 tokens and longer than 70 tokens, as they resulted in difficult relevance annotations. We only kept test dialogues for which at least one sentence was judged as relevant.

We noticed that sentences retrieved for dialogues whose target turn is a Wikipedia grounded turn (i.e., it includes a reference to a Wikipedia section) were somewhat more likely to be judged as relevant than those retrieved for dialogues whose target turn was not Wikipedia grounded. See Table 1 for details. Following, we denoted dialogues with Wikipedia grounded target turns as *Wikipedia grounded dialogues*, and balanced our test set between such dialogues and ungrounded dialogues. As a result, the test set includes 846 dialogues with relevance-judged sentences: 400 ungrounded dialogues and 446 Wikipedia grounded dialogues.

We found that using the thread's subreddit name and title (if exist) is highly important for the initial retrieval method described in Section 3.3, since many times the first turn might be missing or very redundant since the user provided the needed information in the thread title. For example, 202 and 155 dialogues out of 446 Wikipedia grounded dialogues and 400 ungrounded dialogues, have an empty first turn.

Figure 2 shows that most of the test dialogues have a history of 3 turns. Dialogues with a history of 5 turns have more relevant sentences in the initially retrieved list than dialogues with 3 turns in the history, as shown in Figure 3. Dialogues with with history longer than 5 turns are not considered in Figure 3 since the test set includes only a few of them (see Figure 2).

[3]https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

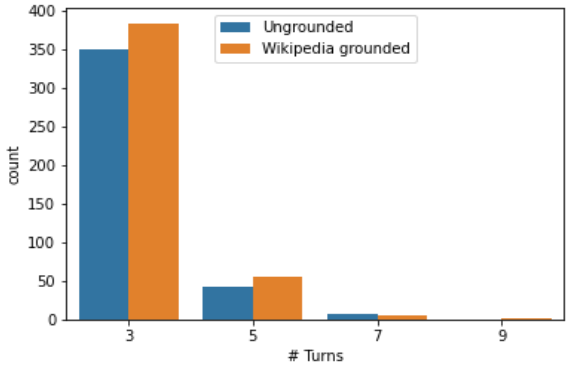


Figure 2: Dialogue count breakdown by the number of turns in the dialogue history (excluding the target turn).

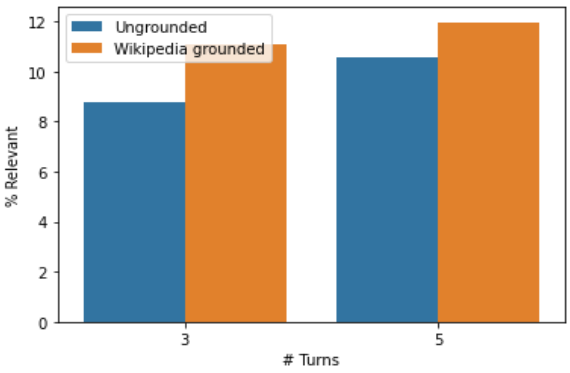


Figure 3: Percentage of relevant candidate sentences for dialogues with a history of 3 or 5 turns.

3.3 Initial Sentence Retrieval

Let g denote a dialogue composed of the turns: $t_1, \ldots, t_n$; t_1 is the first turn and t_n is the last turn. To retrieve sentences for g, we follow common practice in work on passage retrieval [5, 27] and first retrieve documents. We then rank their sentences. The retrieval methods we present are unsupervised; they utilize unigram language models [6].

We start by describing notation. We use $p_x^{MLE}(w)$ to denote the maximum likelihood estimate (MLE) of term w with respect to the text, or text collection, x. $p_x^{Dir}(w)$ is the probability assigned to w by a Dirichlet smoothed unigram language model induced from x [51]. We compare two (unigram) language models, θ_1 and θ_2, using the cross entropy:

$$CE(p(\cdot|\theta_1) \,||\, p(\cdot|\theta_2)) \stackrel{def}{=} -\sum_w p(w|\theta_1) \log p(w|\theta_2).$$

3.3.1 Retrieval Method. For document retrieval, we represent the dialogue g as a linear mixture of language models induced from its turns:

$$p(w|g^{Doc}) \stackrel{def}{=} (1-\beta) p_{t_1}^{MLE}(w) + \frac{\beta}{n-1} \sum_{i=2}^{n} p_{t_i}^{MLE}(w); \quad (1)$$

Table 2: Number of words and the Initial Ranker retrieval score for candidate relevant and non-relevant sentences in the initially retrieved list.

Dialogue type	Label	# Words	Initial Ranker score
Ungrounded	Non-relevant	30	0.53
Ungrounded	Relevant	28.83	0.57
Wikipedia grounded	Non-relevant	31.7	0.55
Wikipedia grounded	Relevant	31.37	0.59

β is a free parameter. Since the first turn, t_1, also contains the dialogue title and the subreddit, we assign it a specific weight. The document corpus is Wikipedia. A Wikipedia document, d, is scored with respect to g using:

$$Score(d; g) \stackrel{def}{=} -CE(p(\cdot|g^{Doc}) \,||\, p_d^{Dir}(\cdot)).$$

The outcome of the document retrieval step is $\mathcal{D}$: the set of top-k retrieved documents.

The next step is to rank the sentences s in $\mathcal{S}$: the set of all sentences of documents in $\mathcal{D}$. For sentence retrieval, we represent the dialogue using a mixture model again. But, in contrast to Equation 1, the emphasis now is on the last turn, t_n:

$$p(w|g^{Sent}) \stackrel{def}{=} (1-\beta) p_{t_n}^{MLE}(w) + \beta \sum_{i=1}^{n-1} \alpha_i p_{t_i}^{MLE}(w); \quad (2)$$

$\alpha_i \stackrel{def}{=} \frac{\delta e^{-\delta|T-i|}}{\sum_{j \in \mathcal{I}} \delta e^{-\delta|T-j|}}$; $T = n-1$, $\mathcal{I} = \{1, \ldots, n-1\}$ and δ is a free parameter (cf., time-based language models [19]). The rationale is that the next turn to be generated for the dialog should be, to some extent, a response to the last turn t_n. The preceding turns serve as the dialogue context and their induced language models are weighed using an exponential decay function. The direct retrieval score of s is defined as:

$$Score(s; g) \stackrel{def}{=} -CE(p(\cdot|g^{Sent}) \,||\, p_s^{Dir}(\cdot)).$$

Finally, following common practice in work on sentence retrieval [27], we integrate the direct retrieval score of s with the retrieval score of its ambient document d_s:

$$FinalScore(s; g) \stackrel{def}{=} (1-\gamma) Score'(d_s; g) + \gamma Score'(s; g), \quad (3)$$

where $Score'(d_s; g)$ and $Score'(s; g)$ are the min-max normalized $Score(d_s; g)$ and $Score(s; g)$ with respect to $\mathcal{D}$ and $\mathcal{S}$, respectively; γ is a free parameter.

Table 2 shows that candidate sentences retrieved for *Wikipedia grounded* dialogues contain more words on average than sentences retrieved for ungrounded dialogues. Relevant (non-relevant) sentences retrieved for grounded dialogues receive higher retrieval scores by the Initial Ranker than relevant (non-relevant) sentences retrieved for ungrounded dialogues. For both grounded and ungrounded dialogues, the average retrieval score for a relevant sentence is higher than that for a non-relevant sentence.

Figure 4 presents the percentage of relevant sentences in the initially retrieved list as a function of the number of terms in the last turn of the dialogue (t_n). We see an overall increasing trend with the main exception being very long turns for ungrounded dialogues. The upward trend can be explained as follows. The sentence ranker

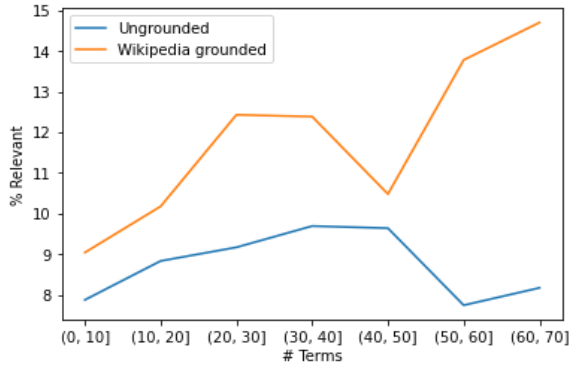


Figure 4: The effect of the number of terms in the last turn (t_n) in the dialogue on the percentage of relevant sentences in the initially retrieved list.

puts emphasis on the last turn (t_n). With more information in this turn, the higher the likelihood to retrieve sentences that contain useful information to produce the next response. These are the sentences that are deemed relevant.

4 SENTENCE RETRIEVAL MODELS FOR OPEN-ENDED DIALOGUES

To provide some sentence retrieval performance results on our proposed dataset, in addition to those of the initial retrieval method described in Section 3.3, we used a few methods described in Section 4.1. Some of these (e.g., neural-based) require a (large) training set. However, generating a large scale manually annotated training set is a laborious effort. Therefore, we instead propose in Section 4.2 a weakly supervised method for automatic labeling.

4.1 Retrieval Methods

We employ a two step sentence retrieval approach. First, we apply the initial sentence retrieval method described in Section 3.3. Then, we re-rank the top-k retrieved sentences using one of the methods proposed below. In contrast to the ad hoc retrieval task, where the information need is explicitly expressed via a query, in open-ended dialogues we have no explicit expression. One proxy to the presumed information need is the last turn t_n: in many cases, as noted above, the next turn to be generated is a response to t_n. Accordingly, herein we use t_n as a query for (re-)ranking the sentences in the initially retrieved list. Utilizing the dialogue context in the retrieval methods described below, as was the case for the initial sentence retrieval method, is left for future work.

LM. We score sentence s using $-CE(p_{t_n}^{MLE}(\cdot) \,||\, p_s^{Dir}(\cdot))$.

Okapi. We assign s its BM25 retrieval score with respect to t_n [38].

IREP$_{\text{BERT}}$. Inspired by a study of a few BERT [11] architectures for ad hoc document retrieval [30], the Independent Representation method, IREP$_{\text{BERT}}$ in short, uses a pre-trained BERT to produce two vectors: (i) for the query (t_n), and (ii) for the sentence s. Specifically, the inputs to BERT for generating the two vectors are "[CLS] t_n [SEP]" and "[CLS] s [SEP]", respectively. The output vectors, which correspond to contextual top-level representation of the [CLS] token, and which are denoted v^{t_n} and v^s, respectively, are used to compute the retrieval score of s via cosine similarity, $Cosine(v^{t_n}, v^s)$.

RANK$_{\text{BERT}}$X. Following Nogueira et al. [29], the RANK$_{\text{BERT}}$X method is a BERT model which is fine-tuned on a retrieval task using dataset X. RANK$_{\text{BERT}}$X produces a relevance score for sentence s with respect to the query t_n through a softmax layer. It gets "[CLS] t_n [SEP] s [SEP]" as an input, and outputs a vector v_{out}, which corresponds to the contextual top-layer representation of [CLS]. The output v_{out} is then fed into a softmax function to estimate the relevance score for s:

$$Score(s; t_n) = Softmax(W_{score}\, v_{out} + b_{score}); \quad (4)$$

W_{score} and b_{score} are learned parameters.

4.2 A Weakly Supervised Training Set

As discussed in Section 2.2, a common practice for training neural retrieval models is to further fine-tune a ranking model for the end task. Since obtaining large scale manually annotated training set is expensive and time consuming, we use weak supervision instead. To this end, we propose an automated method for generating pseudo relevance labels for sentences with respect to conversations created from Reddit. The constructed training set is used in our experiments to fine-tune the RANK$_{\text{BERT}}$ (X) neural retrieval model in a weakly supervised manner.

As training data, we considered every thread that includes Wikipedia grounded turns, each considered as a *target turn*, within the training part of our Reddit collection (see Section 3.1), resulting in 766,334 threads. We filtered out threads with target turns of 5 words or shorter, which we found difficult to automatically label. For each training example (thread), the initial sentence retrieval method (Section 3.3) is used to retrieve 1000 sentences. We only retained threads for which at least one retrieved sentence appears in the Wikipedia sections linked by the Wikipedia grounded target turn. Last, we observed that in threads in which the target turn is followed by more turns in the thread, which we refer to as *future turns*, the Wikipedia section linked in the target turn was of more use within the target turn itself. We therefore kept only threads in which the target turn is followed by additional turns in the thread, ending with 72, 953 threads for training. We refer to these threads as *Reddit conversations*.

Next, each candidate sentence is assigned a pseudo relevance label using a weak annotator. To this end, we rely on a real example of natural continuation of the thread: the target turn and the Wikipedia sections that the author indicated as helpful in writing it. If the turn's author referenced a specific Wikipedia section but not the entire document, we view this as an indication that other paragraphs are less relevant. We note that this assumption does not hold with high confidence for other documents not referenced by the turn's author, because the author might not be familiar with them. Therefore, our weak annotator labels only sentences in the documents containing the pointed sections.

Unlike inference time, where only the conversation history is available, during training set construction the target turn and the future turns are also accessible and can be used to help the pseudo relevance labeling process. Let g of length m denote a conversation

composed of three parts: (a) the conversation history $t_1 \ldots, t_n$, (b) the target turn t_{n+1}, and (c) the future turns $t_{n+2} \ldots, t_m$. To induce pseudo relevance labels using these turns, we used Reciprocal Rank Fusion (RRF) [4] to fuse the scores assigned by four retrieval methods:

Term-based cosine similarity. Cosine similarity between the TF-IDF vectors of a sentence and the target turn t_{n+1}. We use the RSJ IDF version [37] with stopword removal and Krovetz stemming.

BERT-based cosine similarity. Similarly to $\text{IREP}_{\text{BERT}}$ (see Section 4.1), we use a pre-trained BERT [11] to separately embed the candidate sentence s to v^s and the target turn t_{n+1} to $v^{t_{n+1}}$. We then compute the cosine similarity $Cosine(v^{t_{n+1}}, v^s)$.

Fused LM. Let $p_h^{MIX}(w) \stackrel{def}{=} \sum_{i=1}^{n} \alpha_i p_{t_i}^{MLE}(w)$ denote the probability assigned to w by a mixture of language models induced from the turns in the conversation *history*. The language models are increasingly weighted with the exponential decay function α_i from Section 3.3.1 where $T = n$ and $\mathcal{I} = \{1, ..., n\}$. Similarly, let $p_f^{MIX}(w) \stackrel{def}{=} \sum_{i=n+2}^{m} \alpha_i p_{t_i}^{MLE}(w)$ denote the probability assigned to w by a mixture of language models induced from *future* turns, where $T = n+2$ and $\mathcal{I} = \{n+2, ..., m\}$ for the α_i's. We score a sentence s using $-CE(q(\cdot) \mid\mid p_s^{Dir}(\cdot))$, where $q(\cdot) \in \{p_h^{MIX}(\cdot), p_f^{MIX}(\cdot), p_{t_{n+1}}^{MLE}(\cdot)\}$, resulting in the ranked lists L_h, L_f and $L_{t_{n+1}}$. Let $rank_{L_j}(s)$ be the rank of sentence s in L_j; the highest rank is 1. The final score of s is:

$$Score(s) = \frac{\lambda}{2} \frac{1}{\nu + rank_{L_h}(s)} + (1-\lambda) \frac{1}{\nu + rank_{L_{t_{n+1}}}(s)} + \frac{\lambda}{2} \frac{1}{\nu + rank_{L_f}(s)}, \quad (5)$$

where λ and ν are free parameters.

Fused BERT. We utilize a fine-tuned $\text{RANK}_{\text{BERT}}$ (X) from Section 4.1 to rank the sentences against each of the m turns in conversation g. As a result, we obtain m ranked lists $L_i, i \in \{1, ..., m\}$; $rank_{L_i}(s)$ is the rank of sentence s in L_i. Let L_h denote a list fused from the ranked lists induced by the turns in the conversation history. The fusion score of s is $\sum_{i=1}^{n} \alpha_i \frac{1}{\nu + rank_{L_i}(s)}$, where the α_i's are computed with $T = n$ and $\mathcal{I} = \{1, ..., n\}$. Similarly, we create a list L_f by fusing the ranked lists induced by the future turns ($i \in \{n+2, ..., m\}$) where $T = n + 2$ and $\mathcal{I} = \{n + 2, ..., m\}$ for the α_i's. Finally, we assign each sentence s a score which results from fusing L_h, $L_{t_{n+1}}$ and L_f as in Eq. 5.

Once all sentences in a document with a referred section are assigned a retrieval score, the k sentences in the *pointed section* with the highest retrieval scores are labeled as pseudo relevant; the k sentences in the *document* with the lowest retrieval scores, excluding sentences in the pointed section, are labeled as pseudo non-relevant. This selection of pseudo non-relevant sentences strikes a balance between the fact that the sentences might be topically related to the conversation by the virtue of being part of the same document, and the lower likelihood of their relevance due to being the bottom ranked in the document. Figure 5 exemplifies an outcome of the auto-labeling algorithm.

Figure 5: Training example of a conversation created from Reddit with pseudo relevant and non-relevant sentences.

5 EXPERIMENTS

We evaluated the models presented in Section 4 on our novel testset described in Section 3.2. We next describe the setting and the results of these experiments.

5.1 Experimental Setting

5.1.1 *Data Split and Metrics.*

We randomly split the the test set dialogues (Section 3.2) 50 times into two equal-sized sets: validation set, for hyperparameter tuning, and test set. In each split, the number of grounded and ungrounded dialogues is equal between the validation and test sets.

We measure Mean Average Precision (MAP), NDCG of the 5 highest ranked sentences (NDCG@5) and Mean Reciprocal Rank (MRR) of the highest ranked relevant sentence. The results we report are average and standard deviation over the 50 test sets. Statistically significant performance differences are determined using the two tailed permutation (randomization) test with $10,000$ random permutations and $p \leq 0.05$. The tests are performed with respect to the performance attained for each of the 50 test sets. We use Bonferroni correction for multiple hypothesize testing.

5.1.2 *Initial Ranker.*

As document collection we use the Wikipedia dump from 2020-01-01[4]. Parsing was mainly done by Wikiextractor[5] with a few manual improvements. We indexed with Indri[6].

We use the Initial Ranker (Section 3.3) with Krovetz stemming for documents and dialogues, and stopword[7] removal only for dialogues. We did not include candidates that become empty after stopword removal. We retrieve the top-1000 documents for each dialogue and the top-50 sentences for each retrieved document. We

[4]https://dumps.wikimedia.org/enwiki/20200101/enwiki-20200101-pages-articles-multistream.xml.bz2
[5]https://github.com/attardi/wikiextractor.git
[6]http://www.lemurproject.org/indri
[7]http://www.lemurproject.org/stopwords/stoplist.dft

Table 3: Model performance on the testset. 'i' and 'r' mark statistically significant differences with the Initial Ranker and RANK$_{\text{BERT}}$MS→R, respectively. Boldface marks the best performance in a column.

	MAP	NDCG@5	MRR
Initial Ranker	$.238^{\pm.007}$	$.355^{\pm.012}$	$.353^{\pm.012}$
LM	$.185^{\pm.006}_{ir}$	$.253^{\pm.012}_{ir}$	$.256^{\pm.011}_{ir}$
Okapi	$.185^{\pm.006}_{ir}$	$.259^{\pm.010}_{ir}$	$.258^{\pm.009}_{ir}$
IREP$_{\text{BERT}}$	$.172^{\pm.004}_{ir}$	$.236^{\pm.009}_{ir}$	$.240^{\pm.008}_{ir}$
RANK$_{\text{BERT}}$MS	$.328^{\pm.008}_{ir}$	$.457^{\pm.012}_{ir}$	$.444^{\pm.012}_{ir}$
RANK$_{\text{BERT}}$MS→R	$\mathbf{.345}^{\pm.009}$	$\mathbf{.480}^{\pm.013}$	$\mathbf{.461}^{\pm.012}$

Table 4: MAP performance for ungrounded and Wikipedia grounded dialgogues. 'i' and 'r mark statistically significant differences with the Initial Ranker and RANK$_{\text{BERT}}$MS→R, respectively. Boldface marks the best result in a column.

	Ungrounded	Wikipedia Grounded
Initial Ranker	$.223^{\pm.009}_{r}$	$.252^{\pm.010}_{r}$
LM	$.170^{\pm.009}_{ir}$	$.197^{\pm.009}_{ir}$
Okapi	$.168^{\pm.009}_{ir}$	$.197^{\pm.009}_{ir}$
IREP$_{\text{BERT}}$	$.159^{\pm.008}_{ir}$	$.184^{\pm.008}_{ir}$
RANK$_{\text{BERT}}$MS	$.311^{\pm.012}_{ir}$	$.340^{\pm.012}_{ir}$
RANK$_{\text{BERT}}$MS→R	$\mathbf{.323}^{\pm.014}_{i}$	$\mathbf{.362}^{\pm.012}_{i}$

set β to 0.3 in Eq. 1 and 2, γ to 0.75 in Eq. 3, and the Dirichlet prior was set to 1000 [51]. For computing α_i (Section 3.3.1), δ was set to 0.01, following [41].

5.1.3 Training Settings. In all experiments, BERT-Large [11] was used as the pre-trained BERT model. We experimented with two trained RANK$_{\text{BERT}}$X variants with the same architecture (Section 4.1). The first, denoted RANK$_{\text{BERT}}$MS, takes a pre-trained BERT model and fine tunes it on the passage retrieval task in MS Marco [28]. The input is "[CLS] q [SEP] p [SEP]", where q is the query and p is the passage in the MS Marco dataset. The model is trained with a pointwise classification loss[8] [28]. The second variant, RANK$_{\text{BERT}}$MS→R, is a version of RANK$_{\text{BERT}}$MS further fine-tuned, end-to-end, using our weakly supervised training set (Section 4.2), with input as described in Section 4.1. We note that this two stage fine-tuning of BERT-based rankers was also explored in [18].

5.1.4 Hyper-Parameter Tuning.

All hyper-parameters were tuned on the validation sets with MAP as the optimization criterion. We list the hyper-parameters and their corresponding tested value ranges. The Dirichlet smoothing parameter used in LM, μ, is set to values in $\{1000, 2000\}$ [51]. For Okapi BM25, we used $k_1 \in \{1.2, 2, 4, 8, 12\}$ and $b \in \{0.25, 0.5, 0.75, 1\}$.

We fine-tuned a pre-trained BERT on MS Marco using the Adam optimizer with learning rate $\in \{3e-6, 3e-8\}$ and batch size $\in \{64, 128\}$. As in [29], we trained on 12.8M query-passage pairs from the dataset. We further fine-tuned RANK$_{\text{BERT}}$MS with the same training scheme for 10 epochs on our weakly supervised training set, with the hyper-parameter values mentioned above, to obtain RANK$_{\text{BERT}}$MS→R. The maximum sequence length is 512. All turns were truncated to 70 tokens, which is the maximum length of turns in the test set, affecting less than 0.1% of the training set candidates.

For automatic labeling of the training set (Section 4.2), we used RRF with default parameter ν=60, empirically set m future turns to 4, λ to 0.3 in Eq. 5, and and δ to 0.01, when computing α_i for the *Fused LM* and *Fused BERT* methods. We select the top-3 and bottom-3 sentences to serve as the pseudo relevant and non-relevant sentences, respectively. In BERT-based weak annotators, turns and sentences were truncated to contain 64 and 112 tokens, respectively.

5.2 Results

The performance of all evaluated models on the proposed dataset is presented in Table 3. We see that IREP$_{\text{BERT}}$, LM and Okapi, which only match the last turn t_n with a sentence perform statistically significantly worse than the Initial Ranker that matched the sentence with the entire dialogue history. This is a clear evidence that using the dialogue history is important to effectively represent the information need behind the last turn. This finding echoes a similar insight from prior work on response selection for dialogues [44]. It is surprising to see that pre-trained BERT (IREP$_{\text{BERT}}$) underperforms compared to the token-based language models: LM and Okapi. This indicates the importance of fine-tuning BERT models to ranking tasks.

However, once fine-tuned for a retrieval task, a BERT model statistically significantly outperforms all token-based retrieval methods: compare the performance of RANK$_{\text{BERT}}$X methods to that of the other methods in Table 3. This shows that once the dialogue context is taken into consideration in the initial sentence retrieval, re-ranking can improve results with fine-tuned models even when only the last turn and the sentence are matched. In future work, we plan to investigate whether neural models that consider also the dialogue history can further improve performance as indicated in some prior work on conversational search [15, 50].

Table 3 also shows the performance gain of further fine tuning a model on the specific task at hand. Indeed, while RANK$_{\text{BERT}}$MS outperforms all non-fine-tuned models, the RANK$_{\text{BERT}}$MS→R model, which was further fine-tuned using our weakly supervised training set, improves the performance. This method attains the highest performance with all performance gains over other methods being statistically significant. This finding also demonstrates the effectiveness of our weak annotator and weakly supervised training set, showing that performance can be improved without manual annotation for training.

To offer more insight on the types of dialogues in our testset, we computed the MAP of the tested models only on Wikipedia grounded dialogues and only on ungrounded dialogues (see Section 3.2). The performance results, presented in Table 4, show that all models perform better on the Wikipedia grounded dialogues; yet, the relative performance order of methods for ungrounded and grounded diagloues is the same[9]. Thus, we conclude that Reddit conversations that include references to Wikipedia have structure, and embody information, that allow for more effective retrieval than conversations with no such references. In future work we

[8] Training with pairwise loss showed no improvement.

[9] Results for MRR and NDCG@5 show the same patterns as for MAP, and are omitted as they convey no additional insight

plan to investigate whether different model architectures should be applied to each conversation type.

6 CONCLUSIONS AND FUTURE WORK

We introduced the task of sentence retrieval from a document corpus for open-ended dialogues. The goal is to retrieve sentences that contain information useful for generating the next turn in a given dialogue. Sentences that meet this criterion are deemed relevant to the dialogue.

To evaluate retrieval models for the dialogue-based sentence retrieval task, we created a dataset consisting of 846 Reddit dialogues and candidate retrieved sentences from Wikipedia. The dataset also includes human relevance judgments for each sentence. The dataset is available at: https://github.com/SIGIR-2022/A-Dataset-for-Sentence-Retrieval-for-Open-Ended-Dialogues.git.

We are not aware of other publicly available datasets suitable for the evaluation of *retrieval effectiveness* for open-ended dialogues. A unique characteristics of the task is the fact that there is no explicit statement of an information need in the form of questions or queries. Accordingly, the relevance definition we use for sentences is not based on satisfaction of an information need or on being an answer to a question. Rather, as noted above, it is based on the usefulness of the information included in the sentence for generating the next response in the dialogue.

We evaluated several retrieval models on the the novel dataset, including (unigram) language modeling, probabilistic retrieval (Okapi BM25) and neural rankers. To fine-tune neural rankers to the proposed open-dialogue retrieval task, we presented a weak annotator that automatically assigns pseudo-relevance labels to training set sentences. We show that a neural ranker which was fined-tuned using our weakly supervised training set outperforms all other tested models, including a neural ranker fine-tuned on the MS Marco passage retrieval dataset.

In future work, we would like to devise BERT-based retrieval models that are trained based on weak supervision alone, using a pre-trained BERT, without the need for large annotated training sets like MS Marco. We would also like to ground generative language models with our retrieval models and study the conversations that emerge from such grounding.

ACKNOWLEDGEMENTS

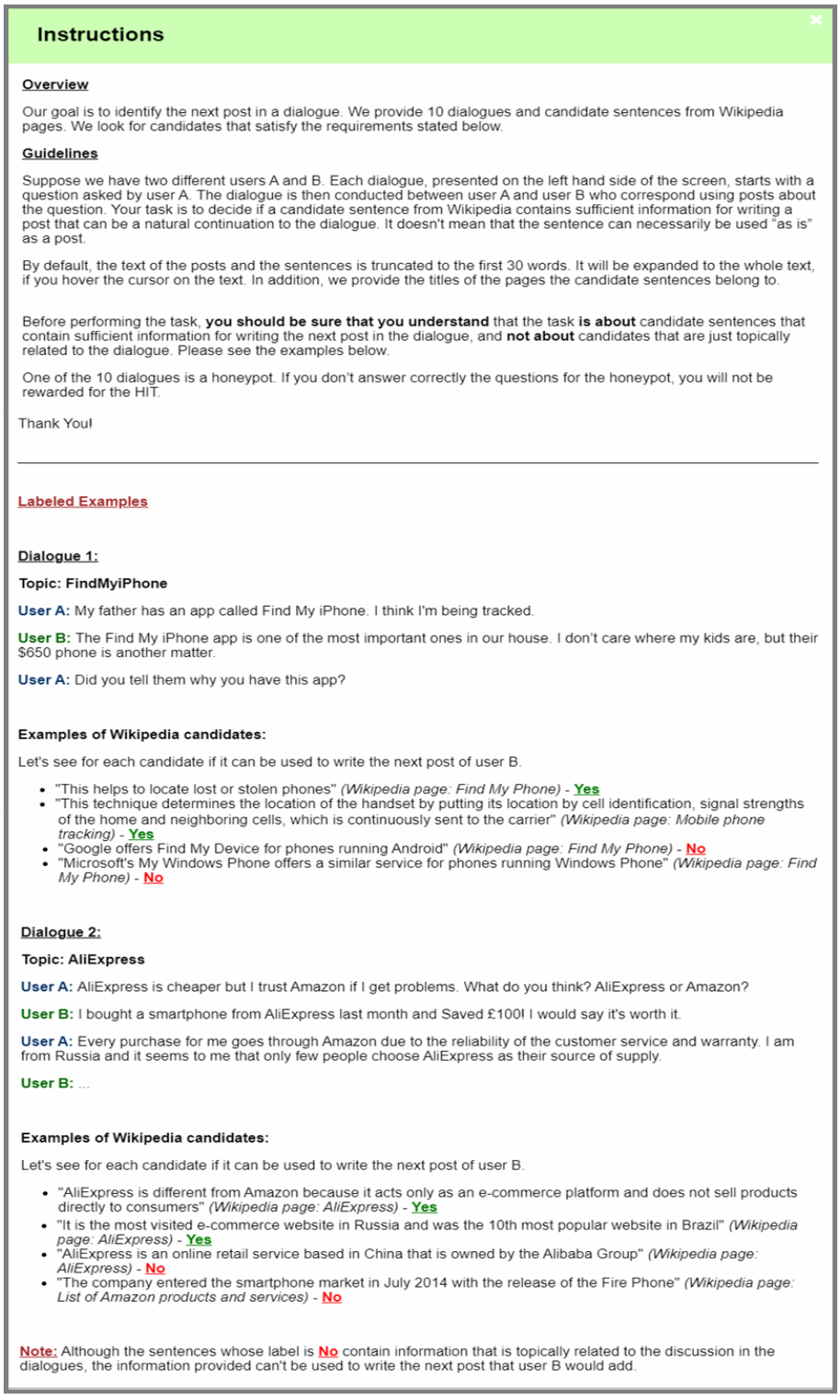
We thank the reviewers for their comments. This work was supported in part by a grant from Google.

A ANNOTATION INSTRUCTIONS

Figure 6 presents the instruction form provided to the annotators.

REFERENCES

[1] Satoshi Akasaki and Nobuhiro Kaji. 2019. Conversation Initiation by Diverse News Contents Introduction. In *Proceedings of NAACL-HLT*. 3988–3998.
[2] Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA - Accessing Domain-Specific FAQs via Conversational QA. In *Proceedings of ACL*. 7302–7314.
[3] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of EMNLP*. 2174–2184.
[4] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion outperforms Condorcet and individual Rank Learning Methods. In *Proceedings of SIGIR*. 758–759.
[5] Andres Corrada-Emmanuel, W. Bruce Croft, and Vanessa Murdock. 2003. *Answer Passage Retrieval for Question Answering*. Technical Report IR-283. Center for Intelligent Information Retrieval, University of Massachusetts.
[6] W. Bruce Croft and John Lafferty (Eds.). 2003. *Language Modeling for Information Retrieval*. Number 13 in Information Retrieval Book Series. Kluwer.
[7] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The Conversational Assistance Track Overview. In *Proceedings of T REC*, Ellen M. Voorhees and Angela Ellis (Eds.), Vol. 1266.
[8] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. *arXiv preprint arXiv:2003.13624* (2020).
[9] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of ACL-CMCL Workshop*. 76–87.
[10] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of SIGIR*. 65–74.
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.
[12] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *Proceedings of ICLR*.
[13] Ahmed Elgohary, Chen Zhao, and Jordan L. Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of EMNLP*. 1077–1083.
[14] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural Approaches to Conversational AI. In *Proceedings of SIGIR*. 1371–1374.
[15] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural Approaches to Conversational Information Retrieval. *CoRR* abs/2201.05176 (2022).
[16] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proceedings of Interspeech*. 1891–1895.
[17] Hsin-Yuan Huang, Eunsol Choi, and Wen tau Yih. 2019. FlowQA: Grasping Flow in History for Conversational Machine Comprehension. In *Proceedings of ICLR*.
[18] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage Representation Aggregation for Document Reranking. *arXiv preprint arXiv:2008.09093* (2020).
[19] Xiaoyan Li and W. Bruce Croft. 2003. Time-Based Language Models. In *Proceedings of CIKM*. 469–475.
[20] Xue Li, Zhipeng Luo, Hao Sun, Jianjin Zhang, Weihao Han, Xianqi Chu, Liangjie Zhang, and Qi Zhang. 2019. Learning Fast Matching Models from Weak Annotations. In *Proceedings of WWW*. 2985–2991.
[21] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. [n.d.]. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of IJCNLP*. 986–995.
[22] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool Publishers.
[23] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Conversational Question Reformulation via Sequence-to-Sequence Architectures and Pretrained Language Models. *arXiv preprint arXiv:2004.01909* (2020).
[24] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *Proceedings of SIGDIAL*. 285–294.
[25] Longxuan Ma, Wei-Nan Zhang, Mingda Li, and Ting Liu. 2020. A Survey of Document Grounded Dialogue Systems (DGDS). *arXiv preprint arXiv:2004.13818* (2020).
[26] Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of EMNLP*. 2322–2332.
[27] Vanessa Murdock. 2007. Aspects of sentence retrieval. *SIGIR Forum* (2007).
[28] 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. (2016).
[29] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).
[30] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. Understanding the Behaviors of BERT in Ranking. *arXiv preprint arXiv:1904.07531* (2019).
[31] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by Reading: Contentful Neural Conversation with On-demand Machine Reading. In *Proceedings of ACL*. 5427–5436.
[32] Chen Qu, Liu Yang, W Bruce Croft, Johanne R Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and Characterizing User Intent in Information-seeking Conversations. In *Proceedings of SIGIR*. 989–992.



Instructions

Overview

Our goal is to identify the next post in a dialogue. We provide 10 dialogues and candidate sentences from Wikipedia pages. We look for candidates that satisfy the requirements stated below.

Guidelines

Suppose we have two different users A and B. Each dialogue, presented on the left hand side of the screen, starts with a question asked by user A. The dialogue is then conducted between user A and user B who correspond using posts about the question. Your task is to decide if a candidate sentence from Wikipedia contains sufficient information for writing a post that can be a natural continuation to the dialogue. It doesn't mean that the sentence can necessarily be used "as is" as a post.

By default, the text of the posts and the sentences is truncated to the first 30 words. It will be expanded to the whole text, if you hover the cursor on the text. In addition, we provide the titles of the pages the candidate sentences belong to.

Before performing the task, **you should be sure that you understand** that the task **is about** candidate sentences that contain sufficient information for writing the next post in the dialogue, and **not about** candidates that are just topically related to the dialogue. Please see the examples below.

One of the 10 dialogues is a honeypot. If you don't answer correctly the questions for the honeypot, you will not be rewarded for the HIT.

Thank You!

Labeled Examples

Dialogue 1:

Topic: FindMyiPhone

User A: My father has an app called Find My iPhone. I think I'm being tracked.

User B: The Find My iPhone app is one of the most important ones in our house. I don't care where my kids are, but their $650 phone is another matter.

User A: Did you tell them why you have this app?

Examples of Wikipedia candidates:

Let's see for each candidate if it can be used to write the next post of user B.

- "This helps to locate lost or stolen phones" *(Wikipedia page: Find My Phone)* - **Yes**
- "This technique determines the location of the handset by putting its location by cell identification, signal strengths of the home and neighboring cells, which is continuously sent to the carrier" *(Wikipedia page: Mobile phone tracking)* - **Yes**
- "Google offers Find My Device for phones running Android" *(Wikipedia page: Find My Phone)* - **No**
- "Microsoft's My Windows Phone offers a similar service for phones running Windows Phone" *(Wikipedia page: Find My Phone)* - **No**

Dialogue 2:

Topic: AliExpress

User A: AliExpress is cheaper but I trust Amazon if I get problems. What do you think? AliExpress or Amazon?

User B: I bought a smartphone from AliExpress last month and Saved £100! I would say it's worth it.

User A: Every purchase for me goes through Amazon due to the reliability of the customer service and warranty. I am from Russia and it seems to me that only few people choose AliExpress as their source of supply.

User B: ...

Examples of Wikipedia candidates:

Let's see for each candidate if it can be used to write the next post of user B.

- "AliExpress is different from Amazon because it acts only as an e-commerce platform and does not sell products directly to consumers" *(Wikipedia page: AliExpress)* - **Yes**
- "It is the most visited e-commerce website in Russia and was the 10th most popular website in Brazil" *(Wikipedia page: AliExpress)* - **Yes**
- "AliExpress is an online retail service based in China that is owned by the Alibaba Group" *(Wikipedia page: AliExpress)* - **No**
- "The company entered the smartphone market in July 2014 with the release of the Fire Phone" *(Wikipedia page: List of Amazon products and services)* - **No**

Note: Although the sentences whose label is **No** contain information that is topically related to the discussion in the dialogues, the information provided can't be used to write the next post that user B would add.

Figure 6: Test set annotation guidelines used in Mechanical Turk.

[33] Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019. BERT with History Answer Embedding for Conversational Question Answering. In *Proceedings SIGIR*. 1133–1136.

[34] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of CHIIR*. 117–126.

[35] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A Conversational Question Answering Challenge. *TACL* 7 (2019), 249–266.

[36] Gary Ren, Xiaochuan Ni, Manish Malik, and Qifa Ke. 2018. Conversational Query Understanding Using Sequence to Sequence Modeling. In *Proceedings of WWW*. 1715–1724.

[37] S.E. Robertson and Walker S. 1997. On relevance weights with little relevance information. In *Proceedings of SIGIR*. 16–24.

[38] Stephen Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Proceedings of TREC-3*. 109–126.

[39] Vasileios Stamatis, Leif Azzopardi, and Alan Wilson. 2019. VES Team at TREC Conversational Assistance Track (CAsT) 2019. In *Proceedings of TREC*.

[40] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query Resolution for Conversational Search with Limited Supervision. In *Proceedings of SIGIR*. 921–930.

[41] Xiaoyan Li W. Bruce Croft. 2003. Time-Based Language Models. In *Proceedings of CIKM*.

[42] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A Dataset for Research on Short-Text Conversations. In *Proceedings EMNLP*. 935–945.

[43] Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning Matching Models with Weak Supervision for Response Selection in Retrieval-based Chatbots. In *Proceedings of ACL*. 420–425.

[44] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of SIGIR*. 55–64.

[45] Rui Yan and Dongyan Zhao. 2018. Coupled Context Modeling for Deep Chit-Chat: Towards Conversations between Human and Computer. In *Proceedings of SIGKDD*. 2574–2583.

[46] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proceedings of EMNLP-IJCNLP*. 3490–3496.

[47] Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S. Lasecki, Jonathan Kummerfeld, Michael Galley, Chris Brockett, Jianfeng Gao, Bill Dolan, Sean Gao, Tim K. Marks, Devi Parikh, and Dhruv Batra. 2018. The 7th dialog system technology challenge. *arXiv preprint* (2018).

[48] Hamed Zamani and W. Bruce Croft. 2018. On the Theory of Weak Supervision for Information Retrieval. In *Proceedings of ICTIR*. 147–154.

[49] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural Query Performance Prediction using Weak Supervision from Multiple Signals. In *Proceedings of SIGIR*. 105–114.

[50] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. *CoRR* abs/2201.08808 (2022).

[51] Chengxiang Zhai and John D. Lafferty. 2001. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of SIGIR*. 268–276.

[52] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of ACL*. 2204–2213.

[53] Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling Multi-turn Conversation with Deep Utterance Aggregation. *arXiv preprint arXiv:1806.09102* (2018).

[54] Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A Dataset for Document Grounded Conversations. In *Proceedings of EMNLP*. 708–713.

[55] Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized Attention-based Deep Network for Conversational Question Answering. *arXiv preprint arXiv:1812.03593* (2018).