Commonsense injection in Conversational Systems: An adaptable framework for query expansion.

Guido Rocchietti
ISTI-CNR, University of Pisa, Pisa, Italy
{guido.rocchietti@isti.cnr.it}

Ophir Frieder
Georgetown University
Washington, DC, USA
ophir@ir.cs.georgetown.edu

Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego
ISTI-CNR, Pisa, Italy
{name.surname@isti.cnr.it}

***Abstract*—Recent advancements in conversational agents are leading a paradigm shift in how people search for their information needs, from text queries to entire spoken conversations.**

This paradigm shift poses a new challenge: a single question may lack the context driven by the entire conversation. We propose and evaluate a framework to deal with multi-turn conversations with the injection of commonsense knowledge. Specifically, we propose a novel approach for conversational search that uses pre-trained large language models and commonsense knowledge bases to enrich queries with relevant concepts. Our framework comprises a generator of candidate concepts related to the context of the conversation and a selector for deciding which candidate concept to add to the current utterance to improve retrieval effectiveness. We use the TREC CAsT datasets and ConceptNet to show that our framework improves retrieval performance by up to 82% in terms of Recall@200 and up to 154% in terms of NDCG@3 as compared to the performance achieved by the original utterances in the conversations.

***Index Terms*—conversational systems, query expansion, commonsense knowledge, KBs, information retrieval**

I. Introduction

Conversational agents, powered by recent advancements in language understanding, are drawing considerable attention. A reason for their success is that they are driving a *paradigm shift* from a "ten blue links" page to a question-answer spoken dialogue interaction, where the assistant provides answers to questions posed by the user. We capitalize on this paradigm shift and propose and evaluate a framework to engineer accurate conversational responses even when context within an utterance is lacking.

This paradigm shift poses interesting challenges. *Utterances* in a conversation, when looked at singularly, may lack the *context* emerging from the entire conversation. Several works in the literature show how to effectively propagate the context of the discussion to each utterance, improving the retrieval performance of the conversational search system. This is achieved by identifying terms previously mentioned in the conversation to profitably expand the current utterance [1]–[4], or by completely rewriting the utterance with a fine-tuned seq2seq neural model [5]–[8]. For instance, QuReTeC [2] is a system based on Bi-LSTM model for query resolution that selects the valuable terms in context to enrich the query. On the other hand, CQR [9] is a rule-based approach to solve coreference and omissions with a finetuned GPT-2 model in a few-shot setting to generate decontextualized queries.

Complementary to these approaches, we exploit commonsense knowledge from external knowledge bases to identify possible concepts that can enrich an utterance and help improve the overall performance of conversational search systems. We claim that using commonsense knowledge is a step forward in the current literature because the state-of-the-art rewriting techniques work at a "term" level, i.e., they enrich the utterance by adding important terms present in the context of the conversation. There are even cases, as shown in our experiments, where relevant *concepts* might not have been previously mentioned in the conversation but can be definitively inferred from the context. These concepts are valuable as they add novel and useful information to the dialogue, e.g., from a sentence like "how do genes work?" we can generate the concept "dna snippet". Our goal is to exploit the capabilities of pre-trained language models and commonsense knowledge bases to improve conversation understanding and enrich the utterance with relevant concepts that may improve the retrieval performance of a conversational search system.

In the past, a significant body of literature showed the importance of correctly identifying and exploiting *entities* [10], i.e., nodes of a network that can be found as fragments of text and that represent a specific person, place, etc. We advance this approach by exploiting commonsense knowledge concepts, i.e., wider entities that represent common things or actions of the world without the need of being named, instead of single terms within the conversational utterance enrichment process.

We first show that using concepts for enriching utterances is effective for the retrieval performance of a conversational search system. Furthermore, we propose a novel framework that exploits the concepts inside a conversation and an external knowledge source to generate commonsense knowledge expansions based on *ConceptNet Numberbatch* [11].

Our framework exploits the concepts inside a conversation and an external knowledge source to generate commonsense knowledge. It includes the following elements:

1) A generator for commonsense knowledge concepts based on Numberbatch embeddings, which uses both query and conversation context;
2) A selector for deciding which candidate concept to add to the current utterance to improve retrieval effectiveness.

We, via an extensive experimental assessment, demonstrate the advantage of the proposed approach on the basis of the results of reproducible experiments conducted with TREC CAsT 2019, 2020, and 2021 datasets. Moreover, we tested the proposed approach with two state-of-the-art rewriting systems confirming the effectiveness also in this setting.

Finally, we performed a reranking phase with an LLM (MonoT5) to observe how our approach impacts the end-to-end performance of a state-of-the-art retrieval system.

The remainder of the paper is organized as follows. In Section II, we overview related work. In Section III, we introduce our commonsense knowledge framework for expanding utterances in conversational search. In Section IV, we present the experimental setup used to validate our proposal. The results of our experiments are presented in Section V. Finally, in Section VI, we conclude and outline some future lines of investigation. To enable reproducibility and foster expansion, our entire framework will be publicly available upon publication.

II. Related Work

a) Conversational Query Rewriting: Conversational information retrieval systems retrieve only a select few of the most relevant documents or passages for a given query while provided with only limited available information found within a conversational query, also referred to as utterance. A commonly used approach to improve accuracy capitalizes on *question rewriting*.

The recent release of ChatGPT, based first on GPT-3.5, and the following release of the GPT-4 [12] that powers both ChatGPT and the Bing search engine poses new challenges in the field, allowing users to perform complex tasks and to submit complex queries. Nonetheless, the size and availability of the models still represent a limit, and we need more feasible ways to treat conversational data.

A proposed way is to address multi-turn dialogues by training a Transformer model that can address co-reference resolution and omissions and rewrite the query by analyzing syntactic relations and references among previous utterances.

Another approach is to create different modules to address different linguistic features in the conversations and then apply a neural re-ranker based on the BERT model to improve the rank of retrieval results [3]. Song et al. [13] train a rewriting neural model on data taken from the intelligent assistant AliMe for the Chinese language after classifying all the words in the data based on their role.

Others propose a rule-based approach to solve co-reference and omissions and then exploit fine-tuned models such as GPT-2 in a few-shot setting to generate a de-contextualized query [9]. Similarly, [7] train a Transformer model using GPT-2 weights to rewrite queries to improve retrieval on TREC CAsT 2019 datasets.

[14] focus on an importance estimation model that evaluates the relevance of every word in each utterance, extracting keywords, measuring utterance ambiguity, and expanding queries with elements selected from previous turns. Finally, they train a Text-To-Text Transfer Transformer (T5) for query rewriting. [1] released a modified version of TREC CAsT 2019, CAsTUR[1], to fine-tune a BERT model to recognize the most critical utterances related to the *current* one, showing promising results.

[4] create a pipeline that classifies utterances based on their references and then enriches them, based on how they were classified, extracting valuable information from the proper context. [15] propose Teresa, a transformer-based model for conversational query rewriting that uses a self-supervised approach to address phenomena such as co-reference and ellipsis. More recently, [16] released Queen, a two-module model that locates all the occurrences of co-reference and ellipsis and then resolves them by rewriting the query with Bert and RoPE [8]. [17] propose an Utterance Rewriting system to apply over chatbots' textual data to improve performances on Natural Language Inference tasks such as contradiction detection by solving ellipsis and anaphoras. More recently, [18]researched a way to exploit LLMs such as ChatGPT to rewrite Utterances. Finally, [19] propose a reinforcement learning approach to improve system capabilities. They use a Conversational Question Answering system as *teacher* and a Query Rewriting model as *student* to increase both the quality of QA and QR models.

b) Knowledge Base for Conversational Search: In the last years, the focus on knowledge bases and their applications in Natural Language Processing tasks has increased. For instance, [20], propose a method to incorporate KB content in a question generation system by training two models, one for predicting relations among entities extracted from the text and the other to generate a tail for a *subject-relation* tuple. The output is fed to a seq2seq model to generate the new question.

c) Our Contribution: The majority of the earlier efforts rely on transformer-based models to rewrite the utterance by addressing phenomena such as ellipsis and anaphora and to exploit the context. In this framework, we propose expanding utterances in conversational search by exploiting commonsense knowledge. Our proposal does not actively rewrite the utterances as other methods in literature do. Thus, as assessed by our experiments, the framework can be effectively used in combination with different approaches for query rewriting to generate even more refined utterances.

III. Commonsense Knowledge for Conversational Search

Our goal is to understand whether we can inject external commonsense knowledge into an utterance to provide the original data with phenomenological information related to the content of the query and the context of the conversation.

We define a multi-turn conversation $\mathcal{U}$ to be a sequence of utterances asked by a user to a conversational assistant, i.e., $u_1, \ldots, u_N$. Let $u_i \in \mathcal{U}$ be the current utterance, while $u_1, \ldots, u_{i-1}$ denote the previous utterances of the same conversation. In the following, we refer to $u_1, \ldots, u_{i-1}$ as *context*. Previous studies show that the context is of paramount importance to deliver high-quality answers as it is a good indicator of the topical evolution of the multi-turn conversation. Given utterance u_i, the goal of a conversational search system is to retrieve a ranked list of documents from a document collection to effectively answer u_i by also keeping track of the context $u_1, \ldots, u_{i-1}$. Our hypothesis is that automatically enriching the utterance with commonsense knowledge from an external source can improve retrieval performance, even in the absence of any other transformation or rewriting of the query.

[1]https://github.com/aliannejadi/castur

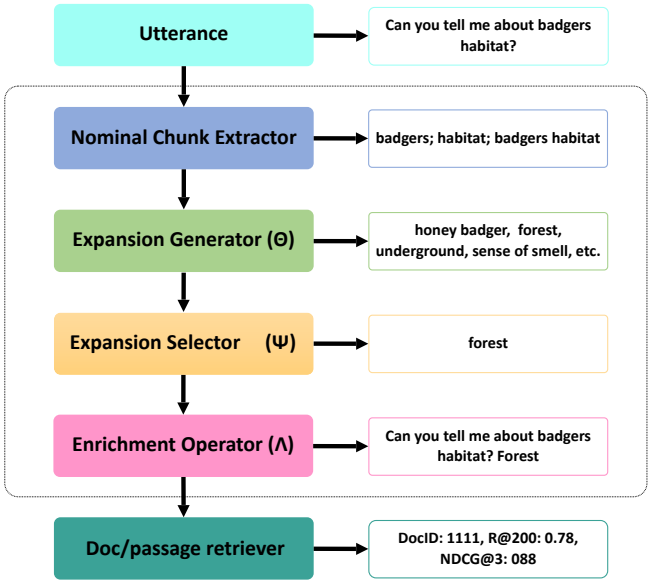

Fig. 1. Architectural diagram of the framework for exploiting commonsense knowledge in conversational search. The described system takes an utterance as input, a nominal chunk extractor selects the nouns, Θ finds the top-k most similar entities, Ψ selects the best one for retrieval, and Λ adds it to the utterance. The retriever uses the new utterance to retrieve a list of results, which is then reordered by the reranker.

To validate our hypothesis, we propose a novel framework composed of three main components:

1) a **commonsense knowledge generator** $\Theta(u_1, \ldots, u_{i-1}, u_i) \mapsto \mathcal{E}^k$ that takes the context and the current utterance to generate a set of k candidate commonsense knowledge expansions, $e_1, \ldots, e_k \in \mathcal{E}$
2) an **expansion selector** $\Psi(u_1, \ldots, u_{i-1}, u_i, e_1, \ldots, e_k) \mapsto \mathcal{E}$ that selects among the k candidate commonsense knowledge expansions the one having the highest likelihood of being useful, $\hat{e} \in \{e_1, \ldots, e_k\}$ given the utterance u_i and its context $u_1, \ldots, u_{i-1}$, in terms of a given quality metric
3) an **utterance enrichment operator** $\Lambda(u_i, \hat{e})$, that takes the current utterance and the selected commonsense knowledge expansion and produces $\hat{u_i}$, i.e., the utterance enriched with $\hat{e}$.

Formally, we propose to inject commonsense knowledge into the conversational search domain by using the combination of the three components below:

$$\hat{u_i} = \Lambda(u_i, \Psi(u_1, \ldots, u_{i-1}, \Theta(u_1, \ldots, u_{i-1}, u_i)))$$

In Figure 1 we show the pipeline using an example. Following, we describe the three components composing our framework.

a) Commonsense Knowledge Generator: The commonsense knowledge generator Θ produces a list of candidate expansions enriching the current utterance u_i given its context $u_1, \ldots, u_{i-1}$. The generation process employs external knowledge bases to overcome the per-term generation inaccuracies exemplified by state-of-the-art query rewriting approaches. While this module is agnostic with respect to the knowledge base employed, we resort to using ConceptNet [2] [11] as the knowledge base for generating external commonsense knowledge. Specifically, Θ generates commonsense knowledge expansions taking as input the nominal chunks extracted from the u_i and its context $u_1, \ldots, u_{i-1}$. In other words, this step looks for entities in the text to derive closely related concepts that can improve the representation of the concept.

Towards this aim, composed noun chunks, e.g., breast cancer, are key in driving an effective expansion generation. For this reason, the nominal chunks identification selects both the composed instances and single ones. The extracted nominal chunks are matched on the knowledge from ConceptNet to see if they are present, and, if so, to select the top k most similar concepts for each one. In Section IV, we describe the implementation details behind the generation of the expansions.

b) Expansion Selector: The expansion selector Ψ is a crucial component of our architecture. Starting from the candidate expansions $e_1, \ldots, e_k \in \mathcal{E}$ obtained by the commonsense knowledge generator Θ, it selects the one with the highest contribution in improving effectiveness when added to the utterance u_i. We consider two different instances of Ψ.

- Ψ_{oracle}, where the expansion selector knows in advance the contribution that each candidate expansion brings to the utterance, and it selects the one with the highest contribution.
- $\Psi_{classifier}$, where the expansion selector does not know in advance the contribution of each candidate expansion. Here, we employ state-of-the-art machine learning techniques to learn classifiers that, given the current utterance u_i, the context $u_1, \ldots, u_{i-1}$, and a candidate expansion e_k, predicts if e_k is an effective expansion for u_i or not.

c) Utterance Enrichment Operator: The utterance enrichment operator Λ is the third component of our architecture. It produces the final enriched utterance $\hat{u_i}$ starting from the selected expansion e_k and the current utterance u_i. Without loss of generality, we define Λ to be the string concatenation operator that produces $\hat{u_i}$ by appending e_k to u_i.

IV. Experimental Setup

We now present the datasets and experimental setup used to validate our commonsense knowledge framework for conversational search.

a) Conversational Datasets: We use the TREC CAsT 2019, 2020, and 2021 datasets with their corresponding evaluation topics and qrels, i.e., judgments made by humans as to whether a document is relevant to an utterance. For 2019, and 2020 TREC CAsT tracks, the collection consists of MS-MARCO [21], TREC CAR [22], and WAPO and has 38,636,520 passages. For 2021, the collection consists of MS-MARCO, KILT [23], and WAPO with 9,679,979 documents in total. The TREC CAsT collections are provided with an evaluation dataset for each year. The datasets are composed of a total of 1,203 utterances divided into 131 conversations. Every conversation is divided into dialogical turns representing only the user part of the conversation, accompanied by an id indicating the document of provenance of the response. Except for the 2019 dataset, they come with three versions of the utterance, the raw utterances, the automatically rewritten, and

[2]https://conceptnet.io/

the manually rewritten (e.g. *raw utterance*: "How do you know when your garage door opener is going bad?"; *manual rewritten utterance*: "How do you know when your garage door opener is going bad?"; *automatic rewritten utterance*: "How do you know when your garage door opener is going bad?"; *manual canonical result id*: "MARCO_5498474").

Finally, qrels files, *"judgments made by humans as to whether a document is relevant to an information need (i.e., topic)"*, enable evaluation.

As relevance judgments are available only for a subset of the conversations provided, we focus our experiments on this subset, composed of 659 utterances in 77 conversations.

b) Commonsense Knowledge Datasets: We select ConceptNet [3] as an external knowledge source, a semantic network in which every node is a concept, e.g., *dog*, *use of devices*, *making toast* etc., and where the concepts are linked through *relations*, e.g., *RelatedTo*, *UsedFor*, *AtLocation*, etc. Specifically, we employ the Numberbatch word embeddings a dataset of conceptual representations of ConceptNet nodes, built using word2vec , GloVe , and OpenSubtitles 2016.

The Numberbatch vocabulary is composed of a total of 516,782 entries. We exploit Numberbatch by extracting the conceptual representation of all the noun chunks present in the conversational utterances and for each noun chunk, the top-25 most similar embeddings.

c) Baselines: We compare our commonsense knowledge framework against two baselines provided by TREC. In fact, TREC CAsT conversations come with:

- *Raw utterances*, informal sentences uttered by a user, representing the real case scenario, often characterized by anaphora or ellipsis, in which the context is implicit rather than explicit.
- *Manual utterances*, well-formed sentences, manually rewritten by humans that explicitly mention the context from the conversation, making questions self-explanatory.

The TREC CAsT 2020 and 2021 datasets are provided with manually rewritten utterances, while for 2019 we use the manually rewritten utterances provided by [4].

Furthermore, we also select two state-of-the-art query rewriting approaches on which we test our framework, QuReTeC [2], a query resolution that rewrites adding terms from the context, and CQR [9] which uses GPT-2 to generate decontextualized queries. Our goal though is not to overcome them but to assess whether our system, at its early stage, can be compared with them.

d) Reproducibility: The code and models used in our experiments are publicly available in our github: https://github.com/hpclab/conv-common-sense.

A. Framework Setup

We now describe how we instantiate the Commonsense Knowledge Framework. We first explain how we exploit Numberbatch within the commonsense knowledge generator θ. We then describe the technical choices behind the expansion selectors presented.

[3]https://github.com/commonsense/conceptnet5/

a) Retrieval settings: We index the TREC CAsT collections using PyTerrier [24] and search them with the DPH weighting model [25]. We evaluate the recall metric computed on the top-200 retrieved documents (R@200) and the Normalized Discounted Cumulative Gain [26] computed on the top-3 results (NDCG@3). We preliminarily preprocess the utterances with a default tokenizer, remove stopwords, and stem tokens using Porter's English stemmer.

b) Commonsense Knowledge Generator: As previously mentioned, we use Numberbatch embeddings as the learned representation of ConceptNet within Θ, to identify the external commonsense knowledge aiming at expanding the raw—unprocessed—utterances. Specifically, we employ SpaCy to process both the current utterance and the context to extract all nominal chunks present in the text. We then search Numberbatch for the vector representation of the extracted concepts and retrieve the top-25 similar ones using cosine similarity.

$$Sim(A, B) = \cos(\alpha) = \frac{A \cdot B}{\|A\|_2 \, \|B\|_2} \tag{1}$$

The generated expansions from Numberbatch are used to evaluate to which extent this new external knowledge overlaps with the knowledge contained in manually-rewritten utterances compared to the raw ones. The main question here is the following: Is Θ able to identify important concepts, namely those manually added in the rewritten version of an utterance? If so, we claim that Numberbatch helps in identifying relevant concepts, important to give the right context to the utterance. We perform this preliminary analysis by extracting the nominal chunks also from the manually-rewritten utterances. By comparing the two sets, we observe an overlap of 72.80%. This confirms that commonsense knowledge helps identify relevant expansions.

The result above is achieved by generating expansions for an utterance given the context, i.e., all the previous utterances of the conversation for which we are generating expansions. We also test the commonsense knowledge generation by employing only the current utterance without the context. In this case, the coverage drops to 39.74%. This result also confirms that context drives the conversation itself.

In the majority of cases, we notice that the extracted expansions are semantically related, and in many cases, they are also constituted by words of different roots (e.g. *semolina* and *granary bread* from *flour)*, concepts that could be beneficial during the retrieval phase, especially for approaches based on bag-of-words.

c) Expansion Selector (Ψ_{oracle}): To evaluate the impact of the generated expansions, we query the index with the raw utterances in the CAsT collections, expanded with each one of the candidates. The results of this preliminary test show that 18.07% of the expanded utterances increase R@200 with respect to the original—raw—utterances, while in only 10.58% of the cases the expansions decrease R@200.

Finally, we assess if the generated expansions are novel knowledge from ConceptNet or if they are concepts already available in the conversation. To this end, we select for each

utterance the expansion maximizing the recall, and we observe that 88.96% of the expansions are novel, i.e., concepts that never appeared in the conversation before. This confirms that our intuition on expanding utterances with commonsense knowledge is a promising idea. Hereinafter, we call Ψ_{oracle} the optimal expansion selector which chooses for each utterance the candidate expansion maximizing the metric of interest, i.e., R@200 or NDCG@3.

d) Expansion Selector ($\Psi_{classifier}$): This expansion selector mimics the behavior of Ψ_{oracle}, via a learned classifier. To train the classifier we label all the expansions improving R@200 (NDCG@3) for the original—raw—baseline as *positive* and all the ones that worsen R@200 (NDCG@3) as *negative*. We train $\Psi_{classifier}$ by using a pre-trained language model, such as BERT [27]. In detail, we train a BERT-based sentence pair classifier that, given the context, the current utterance, and a candidate expansion, predicts to which class the input belongs. To learn the classifier, we introduce two types of input:

- *context and query + expansions* (**CQ+E**): we use the concatenation of the context and the current utterance as the first sentence and the candidate expansion as the second one;
- *context + query and expansion* (**C+QE**): we use the context as the first sentence, while the current utterance and the candidate expansion are concatenated to be the second one.

We learn the sentence pair classifier using the *ktrain* [4] library. We perform k-fold validation with k=5. Furthermore, because the number of *negative* examples in the training set is higher than the *positive* ones, we balance the training data both with one positive, *Balanced* for each negative example and with one positive example for every three negative ones, *Unbalanced*. We also tune the learning rate by performing a grid search to find the optimal parameter value for each dataset. Moreover, we use a validation set randomly generated from the training data during learning to early terminate the training when overfitting is detected.

V. Results and Discussion

Our experiments address the following research questions:

RQ1: Can injecting external commonsense knowledge help a conversational search system?

RQ2: How can we automatically select the best commonsense knowledge concepts to increase retrieval performances?

a) RQ1: Per Section IV-A, we inject commonsense knowledge into user utterances to better specify the meaning of the sentence. The goal is to find whether adding the best concepts to our conversational search system improves effectiveness.

To answer RQ1, we use the previously described oracles and report the results in Table I. As shown in the table our $\Psi_{oracleR@200}$ selector achieves an average R@200 of 0.473 as compared to 0.286 for the original utterances and an average NDCG@3 of 0.295 as compared to 0.162 of the original—raw—utterances, i.e., an 82% relative improvement. For what concerns the $\Psi_{oracleNDCG@3}$ oracle, we observe that the gain for the NDCG@3 obtains an average value of 0.411, interestingly high considering that the raw utterance baseline achieves an NDCG@3 of 0.162 while the manual one 0.400. Without rewriting, merely adding one highly-relevant concept, our oracle almost doubles the results with respect to R@200 and goes even higher if selecting by NDCG@3, increasing by 154% if compared with the raw utterances. It is worth noting that the two improvements achieved by $\Psi_{oracleR@200}$ and $\Psi_{oracleNDCG@3}$ are statistically significant according to the paired t-test with p-value < 0.05.

Thus, we affirm RQ1; adding commonsense knowledge to conversations can improve the retrieval effectiveness of a conversational search system.

b) RQ2: As seen in RQ1, our system can improve retrieval performance up to 154% if it always chooses the best candidate concepts. However, trying all possible combinations to choose the best is not feasible in practice. Thus, we propose $\Psi_{classifier}$, an expansion selector model for predicting the best candidate. All the instances of the classifier are discussed in Section IV-A.

In Table II, we report metrics for each trained classifier computed on the averaged k-fold test sets. We report the accuracy, the F1-score, and the ratios of true positives and negatives. The top four rows present the results for the classifiers trained to maximize R@200; the bottom four rows present the results obtained when maximizing NDCG@3. On four different configurations for each of the selected retrieval metrics (R@200 and NDCG@3), the best-performing models are obtained with a balanced or unbalanced training set depending on the maximizing goal. Furthermore, we report the true positive (TP) and true negative (TN) rates, corresponding to the accuracy calculated by considering only positive and negative examples separately. Regarding accuracy, we achieve 0.893 and 0.862 for CQ+E $_{R@200}$ and CQ+E $_{NDCG@3}$, respectively. We note that the best-performing models in the testing phase are not necessarily the best-performing ones in a real-case scenario; due to the k-fold training and the average of the models' performances.

In Table I, we show the top two models' performances evaluated in a retrieval scenario for MAP@1000, Reciprocal Rank (RR), P@3, P@1, NDCG@3, and R@200. The top two rows are the utterances of the evaluation dataset: Manual and Raw. The following rows are the results obtained with the trained classifiers maximizing R@200 and NDCG@3; the bottom two rows illustrate the results obtained by the two oracles. We use the four models in inference and test the positively predicted expansions to enrich the raw utterances that are then submitted to the retrieval pipeline. As seen, all the best-performing models are trained using both balanced and unbalanced datasets, which did not give us any particular suggestion on the best way of training for our setup. We will further investigate this as part of future work.

In Table III, we report the results obtained evaluating our system in an automatic query rewriting scenario using the CQR and QuReTeC baselines for Mean Average Precision @1000, Reciprocal Rank, P@3, P@1, NDCG@3, and R@200. CQR and QuReTeC are the original baselines; below them, the next four rows, are the two expanded with the oracles,

[4]https://github.com/amaiya/ktrain

TABLE I

RETRIEVAL RESULTS FOR UTTERANCE ENRICHMENT FOR THE TOP-TWO PERFORMING CLASSIFIERSIN BOLD, WE REPORT THE BEST RESULTS ACHIEVED FOR EACH METRIC. WE MARK STATISTICALLY-SIGNIFICANT IMPROVEMENTS (PAIRED t-TEST, p-VALUE < 0.05) WITH RESPECT TO THE PERFORMANCE OF THE RAW UTTERANCES WITH THE SYMBOL ▲.

Type	MAP@1000	RR	P@3	P@1	NDCG@3	R@200
Manual Utterances	**0.299**▲	**0.675**▲	**0.563**▲	0.549▲	0.400▲	**0.598**▲
Raw Utterances	0.123	0.333	0.256	0.225	0.162	0.286
CQ+E$_{R@200}$ *(Balanced)*	0.172▲	0.455▲	0.358▲	0.347▲	0.238▲	0.364▲
CQ+E$_{R@200}$ *(Unbalanced)*	0.161▲	0.419▲	0.314▲	0.306▲	0.197	0.359▲
C+QE$_{R@200}$ *(Balanced)*	0.170▲	0.451▲	0.339▲	0.347▲	0.230▲	0.374▲
C+QE$_{R@200}$ *(Unbalanced)*	0.171▲	0.454▲	0.345▲	0.347▲	0.226▲	0.390▲
CQ+E$_{NDCG@3}$ *(Balanced)*	0.162▲	0.423▲	0.339▲	0.306▲	0.216▲	0.342▲
CQ+E$_{NDCG@3}$ *(Unbalanced)*	0.169▲	0.449▲	0.351▲	0.353▲	0.236▲	0.369▲
C+QE$_{NDCG@3}$ *(Balanced)*	0.161▲	0.450▲	0.345▲	0.347▲	0.235▲	0.340▲
C+QE$_{NDCG@3}$ *(Unbalanced)*	0.137	0.400▲	0.299	0.283	0.194	0.323▲
$\Psi_{oracleR@200}$	0.220▲	0.528▲	0.439▲	0.399▲	0.295▲	0.473▲
$\Psi_{oracleNDCG@3}$	0.195▲	**0.675**▲	0.545▲	**0.601**▲	**0.411**▲	0.370▲

TABLE II

PERFORMANCES OF THE CLASSIFIERS LEARNED FOR EXPANSION SELECTION. IN BOLD, WE HIGHLIGHT THE BEST RESULTS ACHIEVED WITHIN EACH GROUP OF CLASSIFIERS.

Model	Accuracy	F1-score	TP Rate	TN Rate
R@200				
CQ+E *(Balanced)*	**0.893**	**0.892**	0.889	**0.899**
CQ+E *(Unbalanced)*	0.885	0.619	**0.905**	0.644
C+QE *(Balanced)*	0.733	0.530	0.738	0.538
C+QE *(Unbalanced)*	0.843	0.328	0.849	0.496
NDCG@3				
CQ+E *(Balanced)*	0.742	0.704	0.753	0.783
CQ+E *(Unbalanced)*	**0.862**	0.418	**0.877**	0.415
C+QE *(Balanced)*	0.780	**0.754**	0.741	**0.847**
C+QE *(Unbalanced)*	0.838	0.351	0.853	0.379

both NDCG@3 and R@200; at the end, namely the bottom four rows, we present the two baselines expanded with our trained classifiers. We use the queries rewritten by each baseline, released by the authors, and evaluate them. Successively, we apply our framework to both the baselines, testing them with the expansions provided by the oracles and by the best-performing classifier for each objective metric. The classifiers used to expand the baselines are C+QE$_{R@200}$ *(Balanced)* and CQ+E$_{NDCG@3}$ *(Balanced)* for the CQR baseline, and CQ+E$_{R@200}$ *(Unbalanced)* and C+QE$_{NDCG@3}$ *(Balanced)* for QuReTeC. As shown, when considering the oracles, we improve the results of both baselines. Our classifiers overcome the results obtained with the best-performing rewritings of the CQR systems. The most interesting results are the ones regarding $\Psi_{oracleNDCG@3}$, which provided a notable boost to the retrieval results increasing the best-performing baseline ("CQR") by 28% in terms of P@1. Also, we can observe how our framework applied to the CQR baseline, managed to increase most of the metrics in a statistically-significant way, which means that in a first-stage retrieval approach, it can be beneficial to combine the two methods. More interestingly, when combined with the oracle, CQR becomes the overall winning method even if the initial QuReTeC baseline performs better than plain CQR. We also note that CQR + $\Psi_{R@200}$ and QuReTeC + $\Psi_{NDCG@3}$ achieve statistically-significant improvements on two critical metrics for conversational search, i.e., RR and P@1, at the expense of metrics computed at longer cutoffs such as MAP@1000 and R@200, which are less relevant for conversational search, especially after re-ranking. These results suggest that it can be beneficial to expand in such a way to achieve better retrieval results. Additionally, we believe that it would be interesting to develop a similar approach with an end-to-end neural model that could take into account different linguistic features. Likewise, we observe that the distance with the Manual baseline (Table I) is still high in terms of R@200, but we should consider that no rewriting was made during our testing; we added a single entity to the current utterance to significantly increase the retrieval capabilities. Consequently, injecting commonsense external knowledge can boost query "understanding", retrieving the appropriate information and answering RQ2.

c) Reranking with MonoT5: After observing the improvements when adding expansions selected with our framework, we perform a second-phase reranking using the MonoT5 model [28], whose implementation is publicly-available in Pyterrier[5]. In Table IV, we report the results of the reranking phase for the same metrics reported in the previous tables. The top two rows are the results obtained by the original queries of the evaluation dataset; the next two rows present the results obtained by the rewritings of the two baselines. The remaining rows present the results obtained expanding the two baselines with our framework, respectively with the oracles first and then with the best classifiers, both for NDCG@3 and R@200.

We can observe that, except for the QuReTeC baseline expanded with the R@200 best classifier, we managed to increase the performances of the baselines for what concerns P@3, P@1, and NDCG@3. These small cutoff metrics are more representative of the goodness of the system after reranking. For instance, by expanding the CQR baseline with the best NDCG@3 classifier, we managed to increase the P@1 by 4.11%, NDCG@3 by 2.65%, and Reciprocal Rank by 3.34%. Finally, we observe that the baselines expanded using the oracles manage to increase the respective original performance

[5]https://github.com/terrierteam/pyterrier_t5

TABLE III

RESULTS OBTAINED USING OUR FRAMEWORK WITH THE BASELINES REWRITTEN UTTERANCES. IN BOLD, WE REPORT THE BEST RESULTS ACHIEVED FOR EACH METRIC. WE MARK STATISTICALLY-SIGNIFICANT PERFORMANCE GAIN/LOSS (PAIRED t-TEST, p-VALUE < 0.05) OF OUR CORRESPONDING METHODS W.R.T. THE ORIGINAL QURETEC [2] AND CQR [9] BASELINES WITH THE SYMBOLS ▲ AND ▼, RESPECTIVELY.

Type	MAP@1000	RR	P@3	P@1	NDCG@3	R@200
CQR [9]	0.246	0.592	0.482	0.468	0.334	0.520
QuReTeC [2]	0.250	0.625	0.516	0.491	0.349	0.546
CQR+$\Psi_{oracleR@200}$	**0.270**▲	0.648▲	0.538▲	0.520▲	0.366▲	**0.571**▲
CQR+$\Psi_{oracleNDCG@3}$	0.243	**0.702**▲	**0.574**▲	**0.601**▲	**0.419**▲	0.514
QuReTeC+$\Psi_{oracleR@200}$	0.255	0.652	0.536	0.520	0.365	0.552
QuReTeC+$\Psi_{oracleNDCG@3}$	0.238	0.692▲	0.547	0.584▲	0.392▲	0.533
CQR+$\Psi_{R@200}$	0.253	0.642▲	0.507	0.520	0.355	0.542
CQR+$\Psi_{NDCG@3}$	0.237	0.604	0.482	0.486	0.336	0.515
QuReTeC+$\Psi_{R@200}$	0.238▼	0.631	0.497	0.514	0.335	0.525▼
QuReTeC+$\Psi_{NDCG@3}$	0.230▼	0.648	0.495	0.549▲	0.341	0.519▼

TABLE IV

RETRIEVAL RESULTS OBTAINED AFTER PERFORMING RERANKING WITH THE MONOT5 MODEL IN BOLD, WE REPORT THE BEST RESULTS ACHIEVED FOR EACH METRIC. UNDERLINED THE BEST RESULTS ACHIEVED BY OUR MODELS. WE MARK STATISTICALLY-SIGNIFICANT PERFORMANCE GAIN/LOSS (PAIRED t-TEST, p-VALUE < 0.05) OF OUR METHODS WITH RESPECT TO THE ORIGINAL QURETEC [2] BASELINE WITH THE SYMBOLS ▲ AND ▼.

Type	MAP@1000	RR	P@3	P@1	NDCG@3	R@200
Manual Utterances	**0.382**	**0.885**	**0.767**	**0.827**	**0.605**	**0.669**
Raw Utterances	0.173	0.464	0.389	0.399	0.279	0.330
CQR [9]	0.328	0.778	0.676	0.705	0.528	0.591
QuReTeC [2]	0.344	0.785	0.687	0.687	0.533	0.616
CQR + $\Psi_{oracleR@200}$	0.348▲	0.801	0.705	0.717	0.549	0.632▲
CQR + $\Psi_{oracleNDCG@3}$	0.326	0.768	0.686	0.659	0.531	0.593
QuReTeC + $\Psi_{oracleR@200}$	0.347	0.791	0.699	0.694	0.542	0.625
QuReTeC + $\Psi_{oracleNDCG@3}$	0.335	0.788	0.672	0.688	0.519	0.601▼
CQR + $\Psi_{R@200}$	0.330	0.800	0.665	0.728	0.530	0.599
CQR + $\Psi_{NDCG@3}$	0.331	0.804	0.698	0.734	0.542	0.589
QuReTeC + $\Psi_{R@200}$	0.330▼	0.786	0.680	0.688	0.532	0.593▼
QuReTeC + $\Psi_{NDCG@3}$	0.330▼	0.798	0.686	0.711	0.535	0.590▼

for what concerns P@3 and NDCG@3, but in some cases, perform worse than when expanding using our classifier. This suggests that in some cases, the expansion is considered as "noise" by MonoT5. This leaves us with the possibility of further increasing the retrieval metrics and training more complex models for expansion selection and suggests that one should further explore different utterance enrichment operators (Λ). In particular, considering that LLMs are trained on natural language, i.e., in most cases, well-formed sentences, devising a new enrichment operator that better integrates the expansion in the utterance could result in even better results.

VI. CONCLUSIONS AND FUTURE WORK

We proposed a novel way of expanding utterances with external commonsense knowledge to increase the retrieval capabilities of conversational systems. We introduced a theoretical framework exploiting commonsense knowledge in conversational search. Our framework comprises a knowledge generator (Θ), an expansion selector (Ψ), and an utterance enrichment operator (Λ). We discussed the different roles of each framework component and showed how to generate commonsense entities related to conversational utterances using Numberbatch embeddings. We also trained an expansion selector on best and worst-performing expansions in various combinations to produce a total of eight models fine-tuned over a BERT-based transformer model.

Using the TREC CAsT datasets, we demonstrated statistically-significant improvements of up to 82% over the original—raw—utterances when considering R@200 and up to 154% for NDCG@3. We tested our method with two state-of-the-art systems showing a gain in retrieval capabilities for a first-stage retrieval with the DPH weighting model. Finally, we performed a reranking with the MonoT5 model managing to increase the retrieval metrics and observing that the combination of our framework together with state-of-the-art ones can further improve the retrieval phase.

In future work, we want to experiment with mixing other state-of-the-art rewriting techniques with our commonsense knowledge framework. Moreover, some of the limits of our approach come from the fact that the Numberbatch embeddings are not always able to encode text which is not in the vocabulary. These limits can be overcome by using a neural model to encode commonsense knowledge bases. We also intend to contribute in this direction by learning neural representations for commonsense knowledge that allows for building a more effective concept generation. Additionally, we tried using only one generated entity for each utterance. We plan on expanding

the research, to concatenate not only on different entities generated by Numberbatch but also to join different models to provide other knowledge sources.

Finally, the utterance enrichment operator Λ for the presented experiments was a simple concatenation of the selected expansion. We believe that a more sophisticated approach to generating a grammatical sentence expansion could be beneficial for LLMs trained on natural language.

Acknowledgements. Funding for this research has been provided by: PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" funded by the European Union (EU) under the NextGeneration EU programme; the EU's Horizon Europe research and innovation programme EFRA (Grant Agreement Number 101093026). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or European Commission-EU. Neither the EU nor the granting authority can be held responsible for them.

REFERENCES

[1] M. Aliannejadi, M. Chakraborty, E. A. Ríssola, and F. Crestani, "Harnessing evolution of multi-turn conversations for effective answer retrieval," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, ser. CHIIR '20. Association for Computing Machinery, 2020, pp. 33–42. [Online]. Available: https://doi.org/10.1145/3343413.3377968

[2] N. Voskarides, D. Li, P. Ren, E. Kanoulas, and M. de Rijke, "Query resolution for conversational search with limited supervision," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 921–930. [Online]. Available: https://doi.org/10.1145/3397271.3401130

[3] I. Mele, C. I. Muntean, F. M. Nardini, R. Perego, N. Tonellotto, and O. Frieder, "Topic propagation in conversational search," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. ACM, 2020, pp. 2057–2060. [Online]. Available: https://doi.org/10.1145/3397271.3401268

[4] ——, "Adaptive utterance rewriting for conversational search," *Inf. Process. Manag.*, vol. 58, no. 6, p. 102682, 2021. [Online]. Available: https://doi.org/10.1016/j.ipm.2021.102682

[5] S. Yu, J. Liu, J. Yang, C. Xiong, P. Bennett, J. Gao, and Z. Liu, "Few-shot generative conversational query rewriting," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1933–1936. [Online]. Available: https://doi.org/10.1145/3397271.3401323

[6] J. Hao, Y. Liu, X. Fan, S. Gupta, S. Soltan, R. CHADA, P. Natarajan, E. Guo, and G. Tur, "Cgf: Constrained generation framework for query rewriting in conversational ai," in *EMNLP 2022*, 2022.

[7] S. Vakulenko, S. Longpre, Z. Tu, and R. Anantha, "Question rewriting for conversational question answering," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. ACM, 2021, pp. 355–363. [Online]. Available: https://dl.acm.org/doi/10.1145/3437963.3441748

[8] H. Su, X. Shen, R. Zhang, F. Sun, P. Hu, C. Niu, and J. Zhou, "Improving multi-turn dialogue modelling with utterance ReWriter," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 22–31. [Online]. Available: https://aclanthology.org/P19-1003

[9] S. Yu, J. Liu, J. Yang, C. Xiong, P. Bennett, J. Gao, and Z. Liu, "Few-shot generative conversational query rewriting," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. Association for Computing Machinery, 2020, pp. 1933–1936. [Online]. Available: https://doi.org/10.1145/3397271.3401323

[10] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2014.

[11] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: an open multilingual graph of general knowledge," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, pp. 4444–4451.

[12] OpenAI, "Gpt-4 technical report," 2023.

[13] S. Song, C. Wang, Q. Xie, X. Zu, H. Chen, and H. Chen, "A two-stage conversational query rewriting model with multi-task learning," in *Companion Proceedings of the Web Conference 2020*, ser. WWW '20. Association for Computing Machinery, 2020, pp. 6–7. [Online]. Available: https://doi.org/10.1145/3366424.3382671

[14] S.-C. Lin, J.-H. Yang, R. Nogueira, M.-F. Tsai, C.-J. Wang, and J. Lin, "Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting," vol. 39, no. 4, pp. 48:1–48:29, 2021. [Online]. Available: https://doi.org/10.1145/3446426

[15] H. Liu, M. Chen, Y. Wu, X. He, and B. Zhou, "Conversational query rewriting with self-supervised learning," 2021. [Online]. Available: http://arxiv.org/abs/2102.04708

[16] S. Si, S. Zeng, and B. Chang, "Mining clues from incomplete utterance: A query-enhanced network for incomplete utterance rewriting," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022, pp. 4839–4847. [Online]. Available: https://aclanthology.org/2022.naacl-main.356

[17] D. Jin, S. Liu, Y. Liu, and D. Hakkani-Tur, "Improving bot response contradiction detection via utterance rewriting," p. 10, 2022.

[18] E. Galimzhanova, C.-I. Muntean, F. M. Nardini, R. Perego, and G. Rocchietti, "Rewriting conversational utterances with instructed large language models." in *IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology*, 2023.

[19] E. Ishii, B. Wilie, Y. Xu, S. Cahyawijaya, and P. Fung, "Integrating question rewriting in conversational question answering: A reinforcement learning approach," p. 12, 2022.

[20] J. Xin, W. Hao, Y. Dawei, and W. Yunfang, "Enhancing question generation with commonsense knowledge," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*. Chinese Information Processing Society of China, 2021, pp. 976–987. [Online]. Available: https://aclanthology.org/2021.ccl-1.87

[21] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, "MS MARCO: A human generated machine reading comprehension dataset," in *CoCo@ NIPs*.

[22] L. Dietz, M. Verma, F. Radlinski, and N. Craswell, "TREC complex answer retrieval overview," in *Proceedings of {Text REtrieval Conference} (TREC), 2017*., p. 13.

[23] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel, "KILT: a benchmark for knowledge intensive language tasks," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 2523–2544. [Online]. Available: https://aclanthology.org/2021.naacl-main.200

[24] C. Macdonald, N. Tonellotto, S. MacAvaney, and I. Ounis, "PyTerrier: Declarative experimentation in python from BM25 to dense retrieval," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, ser. CIKM '21. Association for Computing Machinery, 2021, pp. 4526–4533. [Online]. Available: https://doi.org/10.1145/3459637.3482013

[25] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '94. Springer-Verlag, 1994, pp. 232–241.

[26] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," vol. 20, no. 4, pp. 422–446, 2002. [Online]. Available: https://doi.org/10.1145/582415.582418

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[28] R. Pradeep, R. Nogueira, and J. J. Lin, "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models," *ArXiv*, vol. abs/2101.05667, 2021.