ISEEQ: Information Seeking Question Generation Using Dynamic Meta-Information Retrieval and Knowledge Graphs

Manas Gaur*[1], Kalpa Gunaratna[2], Vijay Srinivasan[2], Hongxia Jin[2]

[1]AI Institute, University of South Carolina, SC, USA
[2] Samsung Research America, Mountain View, CA, USA
mgaur@email.sc.edu, {k.gunaratna, v.srinivasan, hongxia.jin}@samsung.com

Abstract

Conversational Information Seeking (CIS) is a relatively new research area within conversational AI that attempts to seek information from end-users in order to understand and satisfy users' needs. If realized, such a system has far-reaching benefits in the real world; for example, a CIS system can assist clinicians in pre-screening or triaging patients in healthcare. A key open sub-problem in CIS that remains unaddressed in the literature is generating Information Seeking Questions (ISQs) based on a short initial query from the end-user. To address this open problem, we propose **I**nformation **SEE**king **Q**uestion generator (ISEEQ), a novel approach for generating ISQs from just a short user query, given a large text corpus relevant to the user query. Firstly, ISEEQ uses a knowledge graph to enrich the user query. Secondly, ISEEQ uses the knowledge-enriched query to retrieve relevant context passages to ask coherent ISQs adhering to a conceptual flow. Thirdly, ISEEQ introduces a new deep generative-adversarial reinforcement learning-based approach for generating ISQs. We show that ISEEQ can generate high-quality ISQs to promote the development of CIS agents. ISEEQ significantly outperforms comparable baselines on five ISQ evaluation metrics across four datasets having user queries from diverse domains. Further, we argue that ISEEQ is transferable across domains for generating ISQs, as it shows the acceptable performance when trained and tested on different pairs of domains. The qualitative human evaluation confirms ISEEQ-generated ISQs are comparable in quality to human-generated questions and outperform the best comparable baseline.

Introduction

Information Seeking (IS) is a complex and structured process in human learning that demands lengthy discourse between seekers and providers to meet the information needs of the seekers. The provider can ask the seeker information-seeking questions to understand the seeker's needs better and respond appropriately. For instance, clinicians use their experience or medical knowledge to ask patients information-seeking questions (ISQs), who describe their health condition (a short initial IS query). Conversational Information Seeking (CIS) is an emerging research area within conversational AI that aims to emulate the provider by automatically asking ISQs, keeping track of seeker responses, and ultimately responding to the seeker's needs based on responses to ISQs. CIS has broadened the research scope of various virtual assistants (e.g., Alexa, Bixby) (Zamani and Craswell 2020; Radlinski and Craswell 2017). Existing work in the area of CIS has primarily focused on aspects such as retrieving relevant passages to respond to seeker queries and generating answers (Vakulenko, Kanoulas, and de Rijke 2021; Kumar and Callan 2020).

To the best of our knowledge, the problem of generating ISQs given an initial IS query from the user has not been addressed in the literature so far. Figure 1A shows example ISQs generated for the user IS query "Bothered by feeling down or depressed". For example, user responses to ISQs such as "How often do you feel depressed or hopeless?" and "How long have you struggled with depression?" can be used either by the CIS or the healthcare provider to generate an appropriate response to the user's needs. ISQs differ from other question types (e.g., Clarifying questions, Follow-up questions (Rao and Daumé III 2018) (Zamani et al. 2020) (Pothirattanachaikul et al. 2020)) by having a structure, covering objective details, and expanding on the breadth of the topic. For such a structure between ISQs, there are semantic relations and logical coherence (together termed as *conceptual flow*). From (Aliannejadi et al. 2019), clarifying questions are simple questions of facts, good to clarify the dilemma, and *confined to the entities in the query*. In contrast, ISQs go a step further with expanding the query context by *exploring relationships between entities in the query and linked entities in a knowledge graph*. Thus retrieving a diverse set of passages that would provide a proper solution to a user query.

Firstly, a key challenge in generating ISQs is that the initial IS query is short and has limited context. Without explicit integration of external knowledge for enriching the IS query, CIS cannot achieve the curiosity-driven generation of ISQs (Gaur et al. 2020). Secondly, training an ISQ generation system requires annotated datasets containing IS queries, ISQs, and many passages. Creating such datasets re-

*Research work was done while author was interning at Samsung Research America.

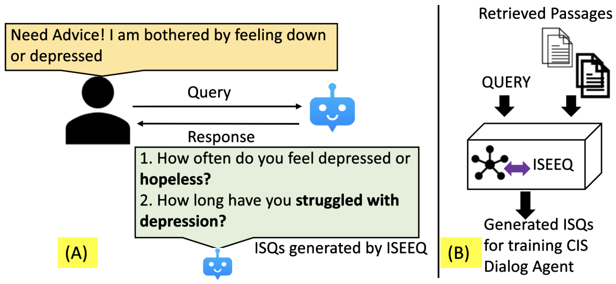


Figure 1: (A) An example of curiosity-driven ISQs generated by ISEEQ. (B) overview of ISEEQ

quires skilled and trained crowdsource workers (Dalton et al. 2020). Moreover, the process is (i) tedious for the crowd worker in terms of the number of passages needed for question creation and can result in (ii) insufficient question coverage when the answer to a query lies across multiple passages, requiring workers to perform extensive search (Wood, Eberhart, and McMillan 2020; Wood et al. 2018).

To address these challenges and the open problem of ISQ generation, we present **I**nformation **SEE**king **Q**uestion generator (ISEEQ) to enable a curiosity-driven CIS system. Essentially, the design of ISEEQ relies on exploring three research questions: **(RQ1)** *Knowledge-infusion*: Can expert-curated knowledge sources like knowledge graphs/bases related to the user query help in context retrieval and question generation? **(RQ2)** *conceptual flow*: Can ISEEQ generate ISQs having semantic relations and logical coherence? **(RQ3)** Can ISEEQ generate ISQs in a cross domain setting and generate ISQs for new domains without requiring crowdsourced data collection?. We believe addressing the three **RQs** uniquely positions this research as the first to develop a successful solution to ISQ generation for CIS. Figure 1B shows the overall inputs and outputs of ISEEQ. ISEEQ generates ISQs based on a short IS query from the seeker, by making use of a large text corpus of passages relevant to the IS query and also relevant knowledge graphs.

Our key contributions of this work are as follows:

1. **Problem definition and approach:** To the best of our knowledge, we are the first to formulate the problem of automatic generation of ISQs for CIS. To solve this, we introduce a novel approach called ISEEQ that can *automatically* generate curiosity-driven and conceptual flow-based ISQs from a short user query.
2. **Dynamic knowledge-aware passage retrieval:** We infuse IS queries with semantic information from knowledge graphs to improve unsupervised passage retrieval. Passages serve as meta-information for generating ISQs.
3. **Reinforcement learning for ISQs:** To improve compositional diversity and legibility in QG, we allow ISEEQ self-guide the generations through reinforcement learning in generative-adversarial setting that results in ISEEQ-RL. We introduce entailment constraints borrowed from natural language inference (NLI) guidelines to expand ISEEQ-RL to ISEEQ-ERL to have smooth topical coherent transitions in the questions, achieving conceptual flow.
4. **Evaluation metrics:** We introduce two evaluation metrics: "semantic relations" and "logical coherence" to measure conceptual flow in the generated questions.

We evaluated ISEEQ (both ISEEQ-RL & ISEEQ-ERL) using four conversational discourse datasets with five natural language generation metrics. In quantitative evaluation, ISEEQ shows superiority over state-of-the-art approaches considered for CIS. We show that ISEEQ is transferable across domains for generating ISQs, as it shows acceptable performance when trained and tested on different pairs of domains; this addresses the key challenge of reducing human effort in training ISQ generation models for new domains. Moreover, 12 human evaluations of 30 IS queries show that ISEEQ generated ISQs are comparable to ground truth human generated questions and they outperformed a competitive baseline generated ones.

Related Work

CIS understands that conversations possess a well-defined structure that addresses the information needs of the user initiating the conversation (Li et al. 2021). The datasets to train models in CIS are designed to facilitate a mixed-initiative dialogue, where the agent can also ask clarifying or follow-up questions to gather information concerning the IS query (Zhang and Bansal 2019). Alongside, datasets like ShARC help CIS models to improve response generation (Saeidi et al. 2018). However, in this study, we restrict ourselves to solving an open sub-problem in CIS that generates ISQs. At present, passage retrieval and ranking, turn-by-turn question generation, search, and answer retrieval have been independently explored within CIS (Vakulenko, Kanoulas, and de Rijke 2021). Dense Passage Retrieval (DPR) retrieves and ranks passages using maximum inner product search (MIPS), using a dense representation based task-specific similarity function. QuAC, HOTPOTQA, WikipassageQA and Topical Chat were introduced to train *IS dialogue* for non-factoid question answering from the retrieved passages (Choi et al. 2018; Yang et al. 2018; Gopalakrishnan et al. 2019; Cohen, Yang, and Croft 2018). Together with DPR, fine-tuning of a transformer language model on either of these datasets can learn better answer retrieval and turn-by-turn question generation, fulfilling part of the requirements for CIS (Lewis et al. 2020).

Despite large-scale crowdsourced annotations, difficulty arises when a CIS agent is required to perform in the presence of a small size IS query. A search through multiple documents to generate logically inferential and semantically related questions for driving mixed-initiative conversations is required (Cho et al. 2021). Also, the agent needs to utilize explicit knowledge to enrich IS queries and broaden the retriever's context to ask questions that would otherwise miss relevant information (Li et al. 2020). The datasets mentioned above lack the characteristics needed in training such a CIS agent. This motivates creating discourse datasets like CAsT-19 (Dalton et al. 2020), used in this research for development and assessment of ISEEQ. Furthermore, there is no study close to ISEEQ that combines explicit knowledge, multi-passage retrieval, and question generation for CIS. Indirectly, ISEEQ gather specific attributes from retrieval augmented generation (RAG) model and retrieval augmented language model (REALM) from (Lewis et al. 2020) and (Guu et al. 2020). However, it significantly improves upon them by (a) preserving the semantic and syntactic structure of the query, (b) use knowledge graphs for passage retrieval, and (c) maintain information flow in question generation. In our evaluations, we utilize RAG components (T5 and DPR) to measure accuracy and quality of ISQs.

Approach

Problem Definition: Given a short query ($q = w_1, w_2, w_3, ..., w_n$) on any topic (e.g., mental health, sports, politics and policy, location, etc.) automatically generate ISQs in a conceptual flow ($ISQ : Q_1, Q_2, Q_3, ..., Q_p$) to understand specificity in information needs of the user.

Our approach to address this problem, ISEEQ, is outlined in Figure 2. We describe in detail the main components of ISEEQ: semantic query expander (SQE), knowledge-aware passage retriever (KPR) and generative-adversarial **R**einforcement **L**earning-based question generator (ISEEQ-RL) with **E**ntailment constraints (ISEEQ-ERL). Inputs to ISEEQ are IS queries described in natural language. For instance, an IS query can be described with **`Titles and Descriptions (T & D)`** (such as in CAsT-19 dataset), **`Descriptions only (D only)`** (such as in QAMR and QADiscourse datasets), **`Topics and Aspects (Tp & Asp)`** (such as in Facebook Curiosity discourse dataset), and others.

SQE: We expand the possibly short user input queries with the help of ConceptNet Commonsense Knowledge Graph (CNetKG)(Speer, Chin, and Havasi 2017). We first extract the entity set $\mathbf{E}_d$ in a user query description d using CNetKG. For this, we use the pre-trained self-attentive encoder-decoder-based constituency parser (Kitaev and Klein 2018) with BERT as the encoder for consistency in ISEEQ. The parser is conditioned to extract noun phrases that capture candidate entities defining an IS query. If the phrases have mentions in the CNetKG they are termed as entities [1]. Then a multi-hop triple (subject-entity, relation, object-entity) extraction over CNetKG is performed using depth first search on entity set $\mathbf{E}_d$. Triples of the form $< e_d, Rel_i, e_x >$ and $< e_y, Rel_j, e_d >$ are extracted where $e_d \in \mathbf{E}_d$. We keep only those triples where e_d ($\in \mathbf{E}_d$) appears as the subject-entity. We use this heuristic (1) to minimize noise and (2) gather more direct information about entities in $\mathbf{E}_d$. Finally, we contextualize d by injecting extracted triples to get k_d, a knowledge augmented query.

Take for example **`D only`** IS query d ($\in \mathbf{D}$), "Want to consider career options from becoming a physician's assistant vs a nurse". The extracted entity set $\mathbf{E}_d$ for d is {career, career_options, physician, physician_assistant, nurse}. Then, the extracted triples for this entity set are <career_options, isrelatedto, career_choice>, <career_options, isrelatedto, profession>, <physician_assistant, is_a, PA>, < physician, is_a, medical doctor>, [...], <nurse, is_a, psychiatric_nurse>, <nurse, is_a, licensed_practical_nurse>, <nurse, is_a, nurse_practitioner>, [...]. The knowledge augmented k_d is "Want to consider career options `career_options is related to career_choice, profession` from becoming a physician's assistant `physician_assistant is_a PA medical doctor, [...]` vs a nurse `nurse is_a psychiatric_nurse, licensed_practical_nurse, [...]` ". Next, we pass this into KPR. The set $\{k_d\}$, $\forall$d $\in \mathbf{D}$ is denoted by $\mathbf{K_D}$ used by QG model in ISEEQ

KPR: Given the knowledge augmented query k_d, KPR retrieve passages from a set $\mathbf{P}$ and rank to get top-K passages $\mathbf{P}_{\text{top-K}}$. For this purpose, we make following specific improvements in the Dense Passage Retriever (DPR) described in (Lewis et al. 2020): (1) Sentence-BERT encoder for the passages $p \in \mathbf{P}$ and k_d. We create dense encodings of $p \in \mathbf{P}$ using Sentence-BERT, which is represented as $\mathcal{Z}_p$ (Reimers and Gurevych 2019). Likewise, encoding of k_d is represented as $\mathcal{Z}_{k_d}$. (2) Incorporate SITQ (Simple locality sensitive hashing (Simple-LSH) and Iterative Quantization) algorithm to pick top-K passages ($\mathbf{P}_{\text{top-K}}$) by using a normalized entity score (NES). SITQ is a fast approximate search algorithm over MIPS to retrieve and rank passages. It can be formalized as $Score(\mathbf{P}_{\text{top-K}}|k_d)$ where,

$$Score(\mathbf{P}_{\text{top-K}}|k_d) \propto \{\text{WMD}(\mathcal{Z}_{k_d}^T \mathcal{Z}_p)\}_{p \in \mathbf{P}}$$

$$\mathcal{Z}_{k_d} = \text{S-BERT}(k_d); \mathcal{Z}_p = \text{S-BERT}(p);$$

SITQ converts dense encodings into low-rank vectors and calculates the semantic similarity between the input query and passage using word mover distance (WMD) (Kusner et al. 2015). $\mathbf{P}_{\text{top-K}}$ from SITQ is re-ranked by NES, calculated[2] for each $p \in \mathbf{P}_{\text{top-K}}$ as $\frac{\sum_{e_j \in k_d} \{\mathbb{I}(e_j = w)\}_{w \in p}}{|k_d|}$ and arrange in descending order. $\mathbf{P}_{\text{top-K}}$ consists of K passages with NES >80%. Execution of KPR is iterative and stopped when each query in the train set has at least one passage for generating ISQs.

We tested retrieving efficiency of KPR using encoding of e_d denoted by $\mathcal{Z}_{e_d}$ and using the encoding of k_d denoted by $\mathcal{Z}_{k_d}$ as inputs to KPR. Measurements were recorded using Hit Rate (HR) @ 10 and 20 retrieved passages. Mean Average Precision (MAP) is calculated with respect to ground truth questions in QAMR. There are two components in MAP: (a) *Relevance* of the retrieved passage in generating questions that have >70% cosine similarity with ground truth; (b) Normalize *Relevance* by the number of ground truth questions per input query. To get MAP, we multiply (a) and (b) and take mean over all the input queries. We computed MAP by setting $K = 20$ retrieved passages due to the good confidence from hit rate (a hyperparameter). KPR outperformed the comparable baselines on the QAMR Wikinews dataset and Table 1 shows that SQE improves the

[1]From here onwards we only use the term Entities, presuming check through exact match is performed using CNetKG

[2]an entity occurring multiple times in p is counted once

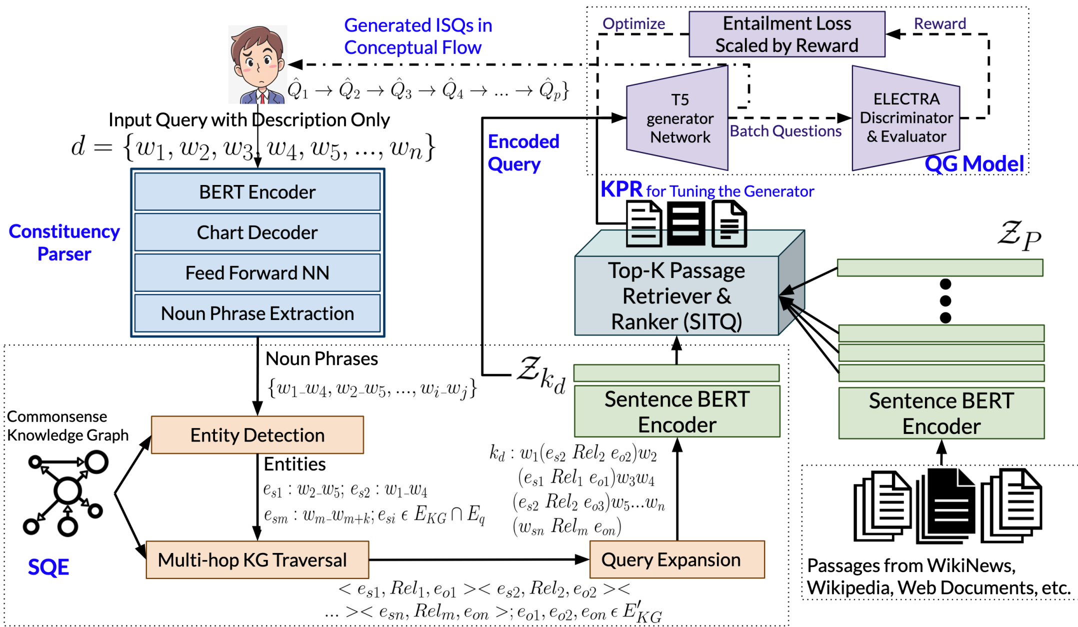


Figure 2: Overview of our approach. ISEEQ combines a BERT-based constituency parser, Semantic Query Expander (SQE), and Knowledge-aware Passage Retriever (KPR) to provide relevant context to a QG model for ISQ generations. The QG Model illustrates a structure of ISEEQ variants: ISEEQ-RL and ISEEQ-ERL. We train ISEEQ in generative-adversarial reinforcement learning setting that maximizes semantic relations and coherence while generating ISQs.

retrieval process[3] A set of $\mathbf{P}_{\text{top-K}}$ for $\mathbf{K_D}$ is denoted by $\{\mathbf{P}_{\text{top-K}}\}_{k_d}$, $k_d \in \mathbf{K_D}$.

QG Model: ISEEQ leverages $\mathbf{K_D}$ and $\{\mathbf{P}_{\text{top-K}}\}_{k_d}$ to learn QG in generative-adversarial setting guided by a reward function. ISEEQ-RL contains T5-base as generator and Electra-base as discriminator to learn to generate IS-type questions. ISEEQ use the reward function to learn to selectively preserve terms from the IS query versus introducing diversity. Also, reward function prevent ISEEQ from generating ISQs that are loose in context or redundant.

Reward Function: Let q_i^n be the i^{th} question in the ground truth questions Q having n tokens and let $\hat{q}_i^m$ be the i^{th} question in the list of generated questions, $\hat{Q}$ having m tokens. We create BERT encodings for each of the n and m words in the question vectors. The reward (R_i) in ISEEQ-RL and ISEEQ-ERL is defined as:

$$\alpha\left[\frac{LCS(\hat{q}_i^m, q_i^n)}{|\hat{q}_i^m|}\right] + (1-\alpha)\left[\sum_{\hat{w_{ij}} \in \hat{q}_i^m} \max_{w_{ik} \in q_i^n} \text{WMD}(\hat{w_{ij}}^T w_{ik})\right] \tag{1}$$

where $\alpha[*]$ is a normalized longest common subsequence (LCS) score that capture word order and make ISEEQ-RL learn to copy in some very complex IS-type queries. $(1-\alpha)[*]$ uses WMD to account for semantic similarity and compositional diversity. For a q_i^n = "What is the average starting salary in the UK?", $(1-\alpha)[*]$ generates $\hat{q}_i^m$="What is the average earnings of nurse in UK?"

Loss Function in ISEEQ-RL: We revise cross entropy (CE) loss for training ISEEQ by scaling with the reward function because each $k_d \in \mathbf{K_D}$ are not only short but they also vary by context. Corresponding to each k_d, there are b ground truth questions $q_{1:b}$ and thus we normalize the revised CE loss by a factor of b. Formally, we define our CE loss in ISEEQ-RL, $\mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta) =$

$$\frac{-\sum_{i=1}^{b} R_i \cdot \mathbb{I}(q_i^n = \hat{q}_i^m) \cdot \mathbf{log} Pr(\hat{q}_i^m|\theta)}{b} \tag{2}$$

where $\mathbb{I}(q_i^n = \hat{q}_i^m)$ is an indicator function counting word indices in q_i^n that match word indices in $\hat{q}_i^m$. The CE loss over $\mathbf{K_D}$ in a discourse dataset is $\mathcal{L}(\hat{Q}|Q, \Theta)_t$, recorded after t^{th} epoch. Formally $\mathcal{L}(\hat{Q}|Q, \Theta)_t$ =

$$\gamma\mathcal{L}(\hat{Q}|Q, \Theta)_{t-1} + (1-\gamma)\mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta) \tag{3}$$

Theoretically, ISEEQ-RL addresses RQ1, but weakly mandates conceptual flow while generating ISQs. Thus, it does not address RQ2.

Loss Function in ISEEQ-ERL: For instance, given d2($\in$ **D**): "Bothered by feeling down or depressed" (shown in Figure 1), ISEEQ-RL generations are: $(\hat{q}_1)$: What is the reason for the depression, hopelessness? and $(\hat{q}_2)$ What is the frequency of you feeling down and depressed? Whereas, ISEEQ-ERL would re-order placing $(\hat{q}_2)$ before $(\hat{q}_1)$ for conceptual flow. To develop ISEEQ-ERL, we redefine the loss function in ISEEQ-RL by introducing principles of entailment as in NLI (Tarunesh, Aditya, and Choudhury 2021)(Gao et al. 2020)[4]. Consider $\hat{q}^m_{i|next}$ to be the

[3] KPR($\mathcal{Z}_{e_d}$) & KPR($\mathcal{Z}_{k_d}$) is executed for each CAsT-19 query.

[4] We use RoBERTa pre-trained on Stanford NLI dataset to measure semantic relations and coherence between a pair of generated questions

next generated question after $\hat{q}_i^m$. We condition equation 2 on $y_{max} = \arg\max_Y \text{RoBERTa}(\hat{q}_i^m, \hat{q}_{i|next}^m)$, where $Y \in \{\text{neutral, contradiction, entailment}\}$ and $Pr(y_{max}) = \max_Y \text{RoBERTa}(\hat{q}_i^m, \hat{q}_{i|next}^m)$. Formally, $\mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta)$ in ISEEQ-ERLis:

if $y_{max} ==$ entailment **then**
  CE $- Pr(y_{max})$
else
  RCE $- (1 - Pr(y_{max}))$
end if

RCE= $-\frac{\sum_{i=1}^{b} R_i(1-\mathbb{I}(q_i=\hat{q}_i))Pr(\hat{q}_i|\theta)}{b}$

Reverse Cross Entropy(RCE) complements CE (Equation 2) by checking $\hat{q}_{i|next}^m$ is semantically related and coherent to $\hat{q}_i^m$. Tuning of the loss after an epoch follows Equation 3.

Datasets

We evaluate ISEEQ-RL and ISEEQ-ERL on a wide range of open-domain knowledge-intensive datasets. Their statistics are shown in Table 2. The datasets exhibit following properties: (1) existence of semantic relations between questions, (2) logical coherence between questions, and (3) diverse context, that is, queries cover wider domains, such as health, sports, history, geography. Fundamentally, these datasets support the assessment of RQ1, RQ2, and RQ3.

QADiscourse (QAD) (Pyatkin et al. 2020) dataset tests the ability of ISEEQ to generate questions that have logical coherence. The sources of queries are Wikinews (WikiN) and Wikipedia (WikiP) that consist of 8.7 Million passages. Question Answer Meaning Representation (QAMR) (Michael et al. 2018) dataset tests the ability of ISEEQ to generate questions with semantic relations between them. The source for creating IS queries is Wikinews, which consist of 3.4 Million passages. Both QAD and QAMR consist of **`D only`** IS queries. Facebook Curiosity (FBC) (Rodriguez et al. 2020) is another dataset that challenges ISEEQ to have both semantic relations and logical coherence. This is because queries are described in the form of **`Tp & Asp`**. The source for IS queries is Wikipedia having 3.3 Million geographical passages. Even though the questions in the dataset have logical coherence, they are relatively less diverse than QAMR and QAD. Conversational Assistance Track (CAsT-19) (Dalton et al. 2020) is the most challenging one for ISEEQ because of size, diversity in context, large number of passages, and IS queries are not annotated with passages. In CAsT-19, IS queries are provided with **`T & D`**.

Retrievers	HR@10	HR@20	MAP
TF-IDF + ECE (Clark et al. 2019)	0.31	0.45	0.16
BM25 + ECE*	0.38	0.49	0.23
DPR (Karpukhin et al. 2020)	0.44	0.61	0.31
KPR($\mathcal{Z}_{e_d}$)	0.47	0.66	0.35
KPR($\mathcal{Z}_{k_d}$)	0.49	0.70	0.38

Table 1: Evaluating retrievers. ECE: Electra Cross Encoder, (*): variant of (Clark et al. 2019), DPR: Dense Passage Retrieval.

Dataset	#Queries(Q/Q)		CNetKG Triples	Source
	Train	Test		
QAD	125 (25)	33 (25)	38.5%	WikiP, WikiN
QAMR	395 (63)	39 (68)	35.5%	WikiN
FBC	8489 (6)	2729 (8)	50%	Geo-WikiP
CAsT-19	30 (9)	50 (10)	57%	MS-MARCO

Table 2: Dataset description. Q/Q: Questions per Query, CNetKG Triples: % of noun/verb phrases identified in CNetKG.

Adapting Datasets: Each dataset, except CAsT-19, has a query, a set of ISQs, and a relevant passage. For fairness in evaluation, we exclude the passages in the datasets; instead, we retrieve them from the sources using KPR. We also perform coreference resolution over ISQs using NeuralCoref to increase entity mentions (Clark and Manning 2016). For example, a question in CAsT-19 "What are the educational requirements required to become one?" is reformulated to "What are the educational requirements required to become a physician's assistant?".

Evaluation and Results

ISEEQ-RL or ISEEQ-ERL generator uses top-p (nucleus) sampling [5] with sum probability of generations equaling to 0.92, a hyperparameter that sufficiently removes the possibility of redundant QG (Holtzman et al. 2019). We evaluate ISEEQ generations using Rouge-L (R-L), BERTScore (BScore) (Zhang et al. 2019), and BLEURT (BRT) (Sellam, Das, and Parikh 2020) that measure preservation of syntactic context, semantics, and legibility of generated question to human understanding, respectively. For conceptual flow in question generation, we define "semantic relations" (SR) and "logical coherence" (LC) metrics. To calculate SR or LC, we pair $\hat{Q}_{1:p}$ generated questions with Q. SR in the generations is computed across all pairs using RoBERTa pretrained on semantic similarity tasks[6]. LC between Q and $\hat{Q}_{1:p}$ is computed from counting the labels predicted as "entailment" by RoBERTa pre-trained on SNLI dataset[7].

Baselines: Since there exists no system to automatically generate ISQs, we considered transformer language models fine-tuned (TLMs-FT) on open domain datasets used for reading comprehension, and complex non-factoid answer retrieval as baselines. Specifically, T5 model fine-tuned (T5-FT) on WikipassageQA (Cohen, Yang, and Croft 2018), SQUAD (Raffel et al. 2019), and CANARD (Lin et al. 2020), and ProphetNet(Qi et al. 2020) fine-tuned on SQUADv2.0 are comparable baselines.

We substantiate our claims in RQ1, RQ2, and RQ3 by highlighting: (1) Multiple passage-based QG yields better results over single gold passage QG used in TLMs-FT (Table 3); (2) Knowledge-infusion through SQE significantly advance the process of QG (Table 3)); (3) Pressing on concep-

[5]Top-p or Top-K sampling either works in ISEEQ
[6]https://huggingface.co/textattack/roberta-base-STS-B
[7]https://paperswithcode.com/lib/allennlp/roberta-snli

Methods	SQE	QAD R-L	QAD BRT	QAD BScore	QAD SR	QAD LC(%)	QAMR R-L	QAMR BRT	QAMR BScore	QAMR SR	QAMR LC(%)	FBC R-L	FBC BRT	FBC BScore	FBC SR	FBC LC(%)
T5-FT WikiPassageQA	-	0.37	0.43	0.16	0.17	10.0	0.19	0.51	0.38	0.36	17.0	0.65	0.78	0.54	0.51	47.3
	+Entities	0.39	0.45	0.16	0.17	10.0	0.20	0.53	0.38	0.36	17.5	0.65	0.78	0.54	0.52	47.4
	+Triples	0.41	0.46	0.16	0.18	11.0	0.20	0.53	0.39	0.37	17.8	0.65	0.78	0.55	0.52	47.3
T5-FT SQUAD	-	0.44	0.54	0.20	0.19	13.0	0.40	0.66	0.46	0.58	21.0	0.70	0.83	0.62	0.67	65.1
	+Entities	0.45	0.56	0.22	0.19	13.5	0.40	0.68	0.47	0.59	22.7	0.71	0.84	0.63	0.69	65.8
	+Triples	0.45	0.58	0.22	0.20	13.8	0.43	0.69	0.47	0.59	22.6	0.70	0.84	0.64	0.69	65.8
T5-FT CANARD	-	0.47	0.54	0.23	0.19	17.1	0.41	0.64	0.53	0.58	22.6	0.73	0.84	0.63	0.67	66.2
	+Entities	0.48	0.55	0.25	0.20	17.5	0.44	0.67	0.62	0.61	23.5	0.74	0.84	0.65	0.69	66.5
	+Triples	0.51	0.57	0.26	0.21	18.3	0.49	0.68	0.66	0.61	24.3	0.74	0.85	0.65	0.70	68.2
ProphetNet-FT SQUAD	-	0.31	0.44	0.14	0.17	12.2	0.35	0.59	0.38	0.36	21.5	0.63	0.78	0.53	0.67	63.2
	+Entities	0.31	0.44	0.14	0.17	12.7	0.37	0.60	0.41	0.37	22.1	0.65	0.78	0.54	0.67	63.3
	+Triples	0.34	0.45	0.15	0.18	13.0	0.37	0.61	0.43	0.37	22.3	0.65	0.79	0.56	0.69	64.0
ISEEQ-RL	-	0.57	0.72	0.40	0.22	20.0	0.50	0.75	0.67	0.64	29.4	0.71	0.84	0.62	0.69	68.2
	+Entities	0.64	0.72	0.41	0.23	22.0	0.52	0.77	0.68	0.64	33.1	0.72	0.85	0.63	0.71	69.8
	+Triples	0.65	0.74	0.45	0.25	22.0	0.53	0.78	0.71	0.65	34.7	0.74	0.87	0.63	0.73	71.8
ISEEQ-ERL	-	0.60	0.76	0.44	0.26	24.5	0.55	0.81	0.72	0.68	36.1	0.74	0.85	0.64	0.76	78.2
	+Entities	0.65	0.78	0.47	0.27	25.2	0.55	0.82	0.74	0.68	36.3	0.77	0.88	0.66	0.76	78.3
	+Triples	0.67	0.79	0.50	0.27	25.7	0.57	0.83	0.77	0.68	37.0	0.79	0.89	0.66	0.78	79.4

Table 3: Scores on test set of datasets. In comparison to T5-FT CANARD, a competitive baseline, ISEEQ-ERL generated better questions across three datasets (30%↑ in QADiscourse, 7%↑ in QAMR, and 5%↑ in FB Curiosity). For fine-tuning we used SQUADv2.0.

tual flow in ISEEQ-ERL improve SR and LC in generations. Evidence from 12 human evaluations support our quantitative findings (Table 6); (4) We investigate the potential of ISEEQ-ERL in minimizing crowd workers for IS dataset creation through cross-domain experiments (Table 5).

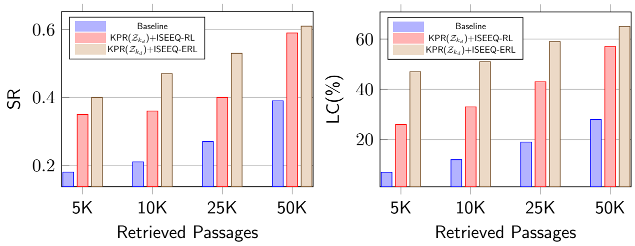
Performance of ISEEQ-RL and ISEEQ-ERL : Datasets used in this research were designed for a CIS system to obtain the capability of multiple contextual passage retrieval and diverse ISQ generations. The process of creating such datasets requires crowd workers to take the role of a CIS system responsible for creating questions and evaluators to see whether questions match the information needs of IS queries. Implicitly, the process embed crowd workers' curiosity-driven search to read multiple passages for generating ISQs. Baselines on employed datasets use single passage QG, with much of the efforts focusing on improving QG. Whereas ISEEQ generation enjoys the success from the connection of SQE, KPR, and novel QG model over baselines in CIS (see Table 3). With SQE, ISEEQ achieved 2-6% across all datasets. The knowledge-infusion in ISEEQ through SQE has shown to be powerful for baselines as well. Table 3 records 3-10%, 3-10%, and 1-3% performance gains of the baselines on QAD, QAMR, and FBC across five evaluation metrics, respectively. SQE allows baselines to semantically widen their search over the gold passages in datasets to generate diverse questions that match better with ground truth. Differently, ISEEQ-RL generations benefit from dynamic meta-information retrieval from multiple passages yielding hike of 20-35%, 6-13%, 3-10% on QAD, QAMR, and FBC, respectively, across five evaluation metrics. Especially, QG in CAsT-19 and FBC datasets advance because of KPR in ISEEQ-RL and ISEEQ-ERL (see Figure 3). Most of the CAsT-19 and FBC queries required multiple passages to construct legible questions. For instance, an IS query : "Enquiry about History, Economy, and Sports in Hyderabad" ISEEQ retrieved following three passages: "History_Hyderabad", "Economy_Hyderabad", and "Sports_Hyderabad" which were missing in the set of passages in FBC. Thus, TLM-FT baselines find it hard to construct legible ISQs using a single passage. Furthermore, ISEEQ-ERL advances the quality of ISQs over ISEEQ-RL by 7-19%, 4-7%, and 5-6% in QAD, QAMR, and FBC (see Table 3) datasets. This is because QAD and FBC questions require the QG model to emphasize conceptual flow.

Ret.Pass.	DPR Train	DPR Test	KPR($\mathcal{Z}_{e_d}$) Train	KPR($\mathcal{Z}_{e_d}$) Test	KPR($\mathcal{Z}_{k_d}$) Train	KPR($\mathcal{Z}_{k_d}$) Test
5K	71	123	99	278	157	275
10K	96	133	154	301	194	316
25K	139	133	235	329	236	363
50K	173	144	269	358	269	402

Table 4: Performance of KPR on MS-MARCO passages while retrieving atleast one passage per IS query in CAsT-19. 269 is the size of CAST-19 train set. KPR covered the train set but left 16% of the IS queries in test set.

Further, we examine the combined **performance of KPR** and ISEEQ-ERL on CAsT-19 dataset. KPR retrieved ∼50K passages sufficient to generate questions for 269 IS queries[8]. Table 4 depicts KPR($\mathcal{Z}_{e_d}$) retrieval performance match KPR($\mathcal{Z}_{k_d}$), with later supported 72% of queries in training set compare to 57% by KPR($\mathcal{Z}_{e_d}$). Also, it outperforms DPR, which supported 30% queries in train set (see Table 4). In test time, KPR($\mathcal{Z}_{k_d}$) supported 84% queries that were used to generate questions by ISEEQ-ERL and evaluated with ground truth for SR and LC (see Figure 3). Apart from monotonic rise in SR and LC scores shown by ISEEQ, ISEEQ-ERL generations achieved better coherence than counterparts with 5K passages (Figure 3 (c) & (d)).

[8]one query can have multiple passages



(a) (b)

Figure 3: Performance improvement of ISEEQ-ERL over ISEEQ-RL and Baseline: T5-FT CANARD concerning SR and LC in generated ISQs. Performed on CAsT-19 with *unannotated* passages.

Test → / Train ↓ (R-L/BRT/BScore/SR/LC(%))	QAD	QAMR	FBC	CAsT-19
QAD	0.67/ 0.79/ 0.50/ 0.27/ 25.7	0.56/ 0.79/ 0.75/ 0.64/ 33.1	0.62/ 0.70/ 0.55/ 0.71/ 73.5	0.76/ 0.48/ 0.64/ 0.60/ 64.2
QAMR	0.73/ 0.89/ 0.62/ 0.28/ 27.7	0.57/ 0.83/ 0.77/ 0.68/ 37.0	0.74/ 0.89/ 0.67/ 0.75/ 77.8	0.67/ 0.41/ 0.57/ 0.57/ 58.6
FBC	0.70/ 0.73/ 0.56/ 0.31/ 33.0	0.61/ 0.85/ 0.72/ 0.67/ 35.8	0.79/ 0.89/ 0.66/ 0.78/ 79.4	0.75/ 0.37/ 0.76/ 0.67/ 66.5
CAsT-19	0.58/ 0.69/ 0.51/ 0.23/ 25.2	0.52/ 0.73/ 0.70/ 0.61/ 33.4	0.63/ 0.77/ 0.57/ 0.73/ 76.5	0.74/ 0.48/ 0.68/ 0.61/ 65.0

Table 5: Transferability test scores using ISEEQ-ERL to answer RQ3. gray cell: ISEEQ-ERL trained and tested on same dataset are along the diagonal. {Train-Test} pairs: {QAD-QAMR}, {CAsT-19-QAMR}, and {QAMR-FBC} showed acceptable cross-domain performance, where train size is smaller than test size.

We attribute the addition of entailment check and RCE for conceptual flow-based QG improvements. **Note:** Qualitative samples of ISQs generated by ISEEQ-ERL, ISEEQ-RL and Baseline (T5-FT CANARD) are provided in github[9] for comparison with ground truth ISQs.

Transferability Test for RQ3: We examine the performance of ISEEQ-ERL in an environment where the train and test dataset belong to a different domain. For instance, QAMR is composed of IS queries from Wikinews, whereas FBC is composed of IS queries from geography category in Wikipedia. From experiments in Table 5, we make two deductions: (1) ISEEQ-ERL provided acceptable performance in generating ISQs for {Train-Test} pairs, where train size is smaller than test size: {QAD-QAMR} and {QAMR-FBC}. (2) ISEEQ-ERL trained on a *narrow domain dataset* (FBC) generated far better ISQs for IS queries across all datasets. The transferability test show ISEEQ-ERL's ability to create new datasets for training and development of CIS systems.

Human Evaluation: We carried out 12 blind evaluations of 30 information-seeking queries covering mental health (7), politics and policy (6), geography (5), general health (3), legal news (2), and others (4). Each evaluator rated ISQs from the ground-truth dataset (S1), ISEEQ-ERL (S2), and T5-FT CANARD (S3) using Likert score where 1 is the lowest and 5 is the highest. Such evaluations takes huge amount of effort; 4 days were invested for high quality evaluations. A total of 570 ISQs (On average 7 by S1, 7 by S2, and 4 by S3) were evaluated on two guidelines, described in Table 6. We measured their statistical significance by first performing one-way ANOVA and then using Least Significant Difference (LSD) post-hoc analysis (Gunaratna et al. 2017). Across the 30 queries on both guidelines, both S1 and S2 are better (statistically significant) than S3 whereas, even though S2 mean is better than S1, there is no statistical significance between the two systems (we may say they are comparable).

	Response: Mean (SD)			$F(2, 957)$	LSD post-hoc
	S1	S2	S3	(p-value)	($p < 0.05$)
G1	3.756 (1.14)	3.759 (1.06)	3.518 (1.08)	5.05 (6.5e-3)	S1>S3, S2>S3
G2	3.803 (1.10)	3.843 (1.02)	3.503 (1.06)	9.71 (6.63e-5)	S1>S3, S2>S3

Table 6: Assessment of human evaluation. G1: ISQs are diverse in context and non-redundant. G2: ISQs are logically coherent and share semantic relations. >: difference is statistically significant. SD: Standard Deviation. S1, S2, and S3 are ground truth, ISEEQ-ERL, and T5-FT CANARD, respectively.

Implementation and Training Details: We implemented our method using Pytorch Lightning on top of the Hugging Face transformer library (Wolf et al. 2019). Hyperparameter tuning in ISEEQ is performed using python library "ray", setting $\alpha = 0.1971$ in equation 1, $\gamma = 0.12$ in equation 3, and learning rate = 1.17e-5. We train ISEEQ for 2 weeks with cross-validation intervals in each epoch, with epochs ranging 100-120 using 4 NVIDIA Tesla V100 GPUs (16GB).

Conclusion

In this research, we introduced, formalized and developed a generic pipeline ISEEQ for generating logically coherent and semantically related ISQs for CIS. ISEEQ outperformed competitive TLM-based baselines in CIS using commonsense knowledge, entailment constraints, and self-guiding through reinforcement learning, trained within a supervised generative-adversarial setting. We established the competency of our method through quantitative experiments and qualitative evaluation on complex discourse datasets. ISEEQ opens up future research directions in CIS by facilitating the automatic creation of large-scale datasets to develop and train improved CIS systems. Crowd-workers can focus on evaluating and augmenting such datasets rather than creating them anew, thus improving dataset standards. The performance of ISEEQ in an online setting was considered beyond the scope of this resource. Broadly construed, through reinforcement learning with the reward on conceptual flow and logical agreement, ISEEQ can be trained to generate questions that are safety constrained and follow a specialized knowledge processing (Sheth et al. 2021).

[9]https://github.com/manasgaur/AAAI-22

References

Aliannejadi, M.; Zamani, H.; Crestani, F.; and Croft, W. B. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *ACM SIGIR*.

Cho, W. S.; Zhang, Y.; Rao, S.; Celikyilmaz, A.; Xiong, C.; Gao, J.; Wang, M.; and Dolan, W. B. 2021. Contrastive Multi-document Question Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 12–30.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Clark, K.; and Manning, C. D. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proc. ACL*.

Cohen, D.; Yang, L.; and Croft, W. B. 2018. WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1165–1168.

Dalton, J.; Xiong, C.; Kumar, V.; and Callan, J. 2020. Cast-19: A dataset for conversational information seeking. In *Proc. of SIGIR*.

Gao, Y.; Wu, C.-S.; Li, J.; Joty, S.; Hoi, S. C.; Xiong, C.; King, I.; and Lyu, M. 2020. Discern: Discourse-Aware Entailment Reasoning Network for Conversational Machine Reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2439–2449.

Gaur, M.; Kursuncu, U.; Sheth, A.; Wickramarachchi, R.; and Yadav, S. 2020. Knowledge-infused deep learning. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 309–310.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tür, D. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*.

Gunaratna, K.; Yazdavar, A. H.; Thirunarayan, K.; Sheth, A.; and Cheng, G. 2017. Relatedness-based multi-entity summarization. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1060–1066.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781.

Kitaev, N.; and Klein, D. 2018. Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2676–2686.

Kumar, V.; and Callan, J. 2020. Making Information Seeking Easier: An Improved Pipeline for Conversational Search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 3971–3980.

Kusner, M.; Sun, Y.; Kolkin, N.; and Weinberger, K. 2015. From word embeddings to document distances. In *International conference on machine learning*, 957–966. PMLR.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Li, S.; Huang, Z.; Cheng, G.; Kharlamov, E.; and Gunaratna, K. 2020. Enriching Documents with Compact, Representative, Relevant Knowledge Graphs. In *IJCAI*, 1748–1754.

Li, Z.; Zhang, J.; Fei, Z.; Feng, Y.; and Zhou, J. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. *arXiv preprint arXiv:2106.02227*.

Lin, S.-C.; Yang, J.-H.; Nogueira, R.; Tsai, M.-F.; Wang, C.-J.; and Lin, J. 2020. Conversational question reformulation via sequence-to-sequence architectures and pretrained language models. *arXiv preprint arXiv:2004.01909*.

Michael, J.; Stanovsky, G.; He, L.; Dagan, I.; and Zettlemoyer, L. 2018. Crowdsourcing Question-Answer Meaning Representations. In *Proc. NAACL*.

Pothirattanachaikul, S.; Yamamoto, T.; Yamamoto, Y.; and Yoshikawa, M. 2020. Analyzing the effects of" people also ask" on search behaviors and beliefs. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, 101–110.

Pyatkin, V.; Klein, A.; Tsarfaty, R.; and Dagan, I. 2020. QADiscourse-Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proc. EMNLP*.

Qi, W.; Yan, Y.; Gong, Y.; Liu, D.; Duan, N.; Chen, J.; Zhang, R.; and Zhou, M. 2020. ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, 2401–2410.

Radlinski, F.; and Craswell, N. 2017. A theoretical framework for conversational search. In *Proc. of CIKM*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rao, S.; and Daumé III, H. 2018. Learning to Ask Good Questions: Ranking Clarification Questions using Neural Expected Value of Perfect Information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2737–2746.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Rodriguez, P.; Crook, P. A.; Moon, S.; and Wang, Z. 2020. Information Seeking in the Spirit of Learning: A Dataset for Conversational Curiosity. In *Proc. of EMNLP*.

Saeidi, M.; Bartolo, M.; Lewis, P.; Singh, S.; Rocktäschel, T.; Sheldon, M.; Bouchard, G.; and Riedel, S. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2087–2097.

Sellam, T.; Das, D.; and Parikh, A. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proc. ACL*.

Sheth, A.; Gaur, M.; Roy, K.; and Faldu, K. 2021. Knowledge-Intensive Language Understanding for Explainable AI. *IEEE Internet Computing*, 25(5): 19–24.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Tarunesh, I.; Aditya, S.; and Choudhury, M. 2021. Trusting RoBERTa over BERT: Insights from CheckListing the Natural Language Inference Task. *arXiv preprint arXiv:2107.07229*.

Vakulenko, S.; Kanoulas, E.; and de Rijke, M. 2021. A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. *arXiv preprint arXiv:2104.07096*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Wood, A.; Eberhart, Z.; and McMillan, C. 2020. Dialogue Act Classification for Virtual Agents for Software Engineers during Debugging. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops*, 462–469.

Wood, A.; Rodeghero, P.; Armaly, A.; and McMillan, C. 2018. Detecting speech act types in developer question/answer conversations during bug repair. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 491–502.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP*.

Zamani, H.; and Craswell, N. 2020. Macaw: An extensible conversational information seeking platform. In *Proc. of ACM SIGIR*.

Zamani, H.; Dumais, S.; Craswell, N.; Bennett, P.; and Lueck, G. 2020. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, 418–428.

Zhang, S.; and Bansal, M. 2019. Addressing Semantic Drift in Question Generation for Semi-Supervised Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2495–2509.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.