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

Manas Gaur  
*University of South Carolina - Columbia*

Kalpa Gunaratna

Vijay Srinivasan

Hongxia Jin

Follow this and additional works at: https://scholarcommons.sc.edu/aii_fac_pub

Part of the Artificial Intelligence and Robotics Commons, and the Data Science Commons

Publication Info
© AAAI, 2022

This Conference Proceeding is brought to you by the Artificial Intelligence Institute at Scholar Commons. It has been accepted for inclusion in Publications by an authorized administrator of Scholar Commons. For more information, please contact dillarda@mailbox.sc.edu.
ISEEQ: Information Seeking Question Generation using Dynamic Meta-Information Retrieval and Knowledge Graphs

Manas Gaur†, Kalpa Gunaratna‡, Vijay Srinivasan†, Hongxia Jin‡
†AI Institute, University of South Carolina, SC, USA
mgaur@email.sc.edu
‡Samsung Research America, Mountain View, CA, USA
{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 the users’ needs. If realized, such a CIS system has far-reaching benefits in the real world; for example, CIS systems can assist clinicians in pre-screening or triaging patients in healthcare. A key open sub-problem in CIS that remains un-addressed 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 Information SEEking Question 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. A qualitative human evaluation confirms that ISEEQ-generated ISQs are comparable in quality to human-generated questions, and it outperformed 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). 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 Daume 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).

Firstly, a key challenge in generating ISQs is that the ini-
tial 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. Secondly, training an ISQ generation system requires annotated datasets containing IS queries, ISQs, and many passages. Creating such datasets requires skilled and trained crowdsourcers \cite{Dalton2020}. 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 \cite{Wood2018}.

To address these challenges and the open problem of ISQ generation, we present Information SEEKing Question 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\footnote{Here on, ISEEQ refers to both ISEEQ-RL & ISEEQ-ERL, unless specified otherwise.} 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 \cite{Li2021}. 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 \cite{Zhang2019}. At present, passage retrieval and ranking, turn-by-turn question generation, search, and answer retrieval have been independently explored within CIS \cite{Vakulenko2021}. Dense Passage Retrieval (DPR) retrieves and ranks passages using maximum inner product search (MIPS), using a dense representation based task-specific similarity function. QuAC, HOT-POTQA, WikipassageQA and Topical Chat were introduced to train IS dialogue for non-factoid question answering from the retrieved passages \cite{Choi2018, Yang2018, Gopalakrishnan2019, Cohen2018}. 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 \cite{Lewis2020}. Despite large-scale crowdsourced annotations, difficulty arises when a CIS model 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\cite{Cho2020}. Also, the model needs to utilize explicit knowledge to enrich IS queries and broaden the retriever’s context to ask questions that would otherwise miss relevant information. The datasets mentioned above lack the characteristics needed in training a CIS agent. This motivates creating discourse datasets like CAsT-19 \cite{Dalton2020}, 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 \cite{Lewis2020} and \cite{Guu2020}. 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 in ISQs.
Approach

Problem Definition: Given a short query \( q = w_1, w_2, w_3, \ldots, w_n \) on any topic (e.g., mental health, sports, politics and policy, location, etc.) automatically generate ISQs in a conceptual flow \( (ISQ: Q_1, Q_2, Q_3, \ldots, Q_p) \) 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 reinforcement learning-based question generator ISEEQ-RL with entailment 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 CAS-T-19 dataset), Descriptions only (D only) (such as in QAMR and QADiscourse datasets), Topics and Aspects (Tp & Asp) (such as in Facebook Curiosity discourses), and others.

SQE: We expand the possibly short user input queries with the help of ConceptNet Commonsense Knowledge Graph (CNetKG)\(^{\text{1}}\) (Speer, Chin, and Havasi 2017). We first extract the entity set \( E_d \) in a user query description \( d \) using CNetKG. For this, we use the pre-trained self-attentive encoder-decoder-based constituency parser \( \text{Kitaev and Klein 2018} \) 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. Then a multi-hop triple (subject-entity, relation, object-entity) extraction over CNetKG is performed using depth first search on entity set \( E_d \). Triples of the form \( < e_d, Rel_i, e_x > \) and \( < e_p, Rel_j, e_d > \) are extracted where \( e_d \in E_d \). We keep only those triples where \( e_d \in E_d \) appears as the subject-entity. We use this heuristic (1) to minimize noise and (2) gather more detailed information about entities in \( E_d \). Finally, we contextualize \( d \) by injecting extracted triples to get \( k_d \), a knowledge augmented query.

Take for example D only IS query \( d \in (D) \), “Want to consider career options from becoming a physician’s assistant vs a nurse”. The extracted entity set \( E_d \) for \( d \) is \{career, career_options, physician, physician_assistant, nurse\}. Then, the extracted triples for this entity set are \(<\text{career_options}, \text{isrelatedto}, \text{career_choice}>, <\text{career_options}, \text{isrelatedto}, \text{profession}>, <\text{physician_assistant, is}a, \text{PA}>, <\text{physician, is}a, \text{medical_doctor}>, [...], <\text{nurse, is}a, \text{psychiatric_nurse}>, <\text{nurse, is}a, \text{licensed_practical_nurse}>, <\text{nurse, is}a, \text{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 \( \{k_d\}, \forall d \in D \) is denoted by \( K_D \).

KPR: Given the knowledge augmented query \( k_d \), KPR retrieve passages from a set \( P \) and rank to get top-K passages \( P_{\text{top-K}} \). For this purpose, we make following specific improvements in the Dense Passage Retriever (DPR) described in \( \text{Lewis et al. 2020} \): (1) Sentence-BERT encoder for the passages \( p \in P \) and \( k_d \). We create dense encodings of \( p \in P \) using Sentence-BERT, which is represented as \( Z_p \)\(^{\text{2}}\) (Reimers and Gurevych 2019). Likewise, encoding of \( k_d \) is represented as \( Z_{k_d} \). (2) Incorporate SITQ (Simple locality sensitive hashing (Simple-LSH) and Iterative Quantization) algorithm to pick top-K passages \( P_{\text{top-K}} \) 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 \( \text{Score}(P_{\text{top-K}} | k_d) \) where,

\[
\text{Score}(P_{\text{top-K}} | k_d) \propto \{\text{WMD}(Z_{k_d}^T Z_p)\}_{p \in P}
\]

\( Z_{k_d} = S\text{-BERT}(k_d); Z_p = S\text{-BERT}(p); \)

SITQ converts dense encodings into low-rank vectors and calculates the semantic similarity between the input query and passage using word mover distance (WMD)\(^{\text{3}}\) (Kusner et al. 2015). \( P_{\text{top-K}} \) from SITQ is re-ranked by NES, calculated\(^{\text{4}}\) for each \( p \in P_{\text{top-K}} \) as \( \sum_{i,j \in E_k} \text{Sim}(e_i, e_p) \) \( |k_d| \) and arrange in descending order. \( P_{\text{top-K}} \) 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 \( Z_{e_d} \) and using the encoding of \( k_d \) denoted by \( Z_{k_d} \) 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 retrieval process\(^{\text{5}}\). A set of \( P_{\text{top-K}} \) for \( K_D \) is denoted by \( \{P_{\text{top-K}} \}_{k_d} \) \( k_d \in K_D \).

QB Model: ISEEQ leverages \( K_D \) and \( \{P_{\text{top-K}} \}_{k_d} \) to learn QB in generative-adversarial setting guided by a reward function. ISEEQ-RL contains TS-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

\(^{\text{1}}\)From here onwards we only use the term Entities, presuming check through exact match is performed using CNetKG

\(^{\text{2}}\)SITQ contains ISEEQ-RL and ISEEQ-ERL

\(^{\text{3}}\)an entity occurring multiple times in p is counted once

\(^{\text{4}}\)KPR(\(Z_{e_d}\)) is executed for all the IS queries in CAS-T-19 independently. Same for KPR(\(Z_{k_d}\))

\(^{\text{5}}\)KPR(\(Z_{e_d}\)) is executed for all the IS queries in CAS-T-19 independently. Same for KPR(\(Z_{k_d}\))
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.

Loss Function in ISEEQ-RL: We revise cross entropy (CE) loss for training ISEEQ by scaling with the reward function because each $k_d \in \mathbb{K}_D$ 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, $\mathcal{L}(\hat{q}_{1:b} | q_{1:b}, \theta) = \sum_{i=1}^{b} R_i : I(\hat{q}_i = \hat{q}_m) \cdot \log Pr(q_i | q_{1:b}, \theta)$.

Loss Function in ISEEQ-ERL: For instance, given $d_2 \in D$: “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 (Harunesh, Aditya, and Choudhury 2021) Consider $\hat{q}_m^{\hat{q}_{1:b}}$ to be the next gen-

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

<table>
<thead>
<tr>
<th>Retrievers</th>
<th>HR@10</th>
<th>HR@20</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF + ECE (Clark et al. 2019)</td>
<td>0.31</td>
<td>0.45</td>
<td>0.16</td>
</tr>
<tr>
<td>BM25 + ECE*</td>
<td>0.38</td>
<td>0.49</td>
<td>0.23</td>
</tr>
<tr>
<td>DPR (Karpukhin et al. 2020)</td>
<td>0.44</td>
<td>0.61</td>
<td>0.31</td>
</tr>
<tr>
<td>KPR($Z_{q_d}$)</td>
<td>0.47</td>
<td>0.66</td>
<td>0.35</td>
</tr>
<tr>
<td>KPR($Z_{k_d}$)</td>
<td>0.49</td>
<td>0.70</td>
<td>0.38</td>
</tr>
</tbody>
</table>

diversity. Also, reward function prevent ISEEQ from generating ISQs that are too dense or redundant.

**Reward Function:** Let $q_i^n$ be the $i^{th}$ question in the ground truth questions $Q$ having $n$ tokens and let $q_i^m$ be the $i^{th}$ question in the list of generated questions, $Q$ 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-E RL is defined as:

$$ R_i = \alpha \left[ \frac{LCS(q_i^n, q_i^m)}{|q_i^m|} \right] + (1- \alpha) \left[ \sum_{w_{i,j} \in q_i^n} \max_{w_{i,k} \in q_i^m} \text{WMD}(w_{i,j}^T, w_{i,k}) \right] $$

(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 $q_i^m$ = “What is the average earnings of nurse in UK?”.

*ISEEQ generated questions are often grammatically correct
erated question after $\hat{q}_m$. We condition equation 2 on $y_{\text{max}} = \arg\max_Y \text{RoBERTa}(q_{m_1}^m, q_{m_1}^{m_1})$, where $Y \in \{\text{neutral, contradiction, entailment}\}$ and $Pr(y_{\text{max}}) = \max_Y \text{RoBERTa}(q_{m_1}^m, q_{m_1}^{m_1})$. Formally, $L(q_{1:2}|q_{1:3}, \theta)$ in ISEEQ-ERL is:

$$
\begin{align*}
\text{if } y_{\text{max}} &= \text{entailment} \text{ then } \\
&CE − Pr(y_{\text{max}}) \\
\text{else} \text{ end if} \\
RCE = −\sum_{i=1}^{b} R_i(1−Pr(q_i(\theta)))Pr(q_i(\theta))
\end{align*}
$$

Reverse Cross Entropy (RCE) complements CE (equation 2) by checking if $q_{m_1}^{m_1}$ is semantically related and coherent to $q_{m_1}$. A caveat in ISEEQ-ERL is addition of epochs, since in initial few epochs the model selects the else over if condition. Tuning of the loss after an epoch follows equation 3.

**Datasets**

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

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

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.

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 with sum probability of generations equal 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 pre-trained on semantic similarity task. LC between $Q$ and $\hat{Q}_{1:p}$ is computed from counting the labels predicted as “entailment” by RoBERTa pre-trained on SNLI dataset.

**Baselines:** Since there exists no system to automatically generate ISQs, we considered transformer language models fine-tuned (TLMs-HT) on open domain datasets used for reading comprehension, and complex non-factoid answer retrieval as baselines. Specifically, T5 model fine-tuned (T5-HT) on WikipediaQA (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-HT (Table 3); (2) Knowledge-infusion through SQE significantly advance the process of QG (Table 4); (3) Pressing on conceptual flow in ISEEQ-ERL improve SR and LC in generations. Evidence from 12 human evaluations support our quantitative findings (Table 7); (4) We investigate the potential of ISEEQ-ERL in minimizing crowd workers for IS dataset creation through cross-domain experiments (Table 6).
Table 3: Scores on test set of datasets. In comparison to T5-FT CANDAR, 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.

Table 4: An ablation study showing improvement in the quality of ISQs after encodings of retrieved passages (P_e,K) are concatenated with knowledge-augmented query (k_d) after SQE. The concatenation is performed for each p ∈ P_e,K.

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). 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). 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 advance the quality of ISQs over ISEEQ-RL by 7-14% and 6-10% in QAD and FBC. This is because QAD and FBC questions require the QG model to emphasize conceptual flow.

Table 5: Performance of KPR on MS-MARCO passages while retrieving at least 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 retrieve ~50K passages sufficient to generate questions for 269 IS queries. Table 5 depicts KPR(ζ_e,d) retrieval performance match KPR(ζ_k), with later supported 72% of queries in training set compare to 57% by KPR(ζ_e,d). Also, it outperforms DPR, which supported 30% queries in train set (see Table 5). In test time, KPR(ζ_e,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 5(c) & (d)). We attribute the addition of entailment check and RCE for conceptual flow-based QG improvements. Note: We provide samples of ISQs under supplementary material. ISQs are generated by ISEEQ-ERL, ISEEQ-RL and Baseline (T5-FT CANARD) for comparison with ground truth ISQs.

**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 rate 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. 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 7. We measured their statistical significance by first performing one-way ANOVA and then using Least Significant Difference (LSD) post-hoc analysis (as performed in 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 on the scores (we may say they are comparable).

**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 with cross-validation intervals in each epoch, with epochs ranging 100-120 depending on the dataset size. Four NVIDIA Tesla V100 GPUs (16GB) were used for two weeks of training.

**Conclusion**

In this research, we introduced, formalized and developed a generic pipeline ISEEQ for generation 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 discourse datasets. Crowd-workers can focus on evaluating and augmenting such datasets rather than creating them anew, thus improving dataset standards.

**Table 7:** 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.

<table>
<thead>
<tr>
<th>Test $\rightarrow$</th>
<th>QAD $\rightarrow$</th>
<th>QAMR $\rightarrow$</th>
<th>FBC $\rightarrow$</th>
<th>CaSt-19 $\rightarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train ↓</td>
<td>R-L/BRT/IScore/SLC(%)</td>
<td>R-L/BRT/IScore/SLC(%)</td>
<td>R-L/BRT/IScore/SLC(%)</td>
<td>R-L/BRT/IScore/SLC(%)</td>
</tr>
<tr>
<td>QAD</td>
<td>0.67/ 0.79/ 0.56/ 0.79/ 0.62/ 0.70/ 0.76/ 0.48/ 0.50/ 0.47/ 0.75/ 0.64/ 0.55/ 0.71/ 0.64/ 0.60/ 25.7</td>
<td>0.57/ 0.63/ 0.62/ 0.68/ 0.67/ 0.75/ 0.57/ 0.57/ 0.67/ 0.67/ 0.75/ 0.75/ 0.66/ 0.66/ 0.78/ 0.78/ 0.76/ 0.67/ 32.7</td>
<td>0.70/ 0.73/ 0.61/ 0.85/ 0.79/ 0.89/ 0.75/ 0.37/ 0.56/ 0.31/ 0.72/ 0.67/ 0.56/ 0.78/ 0.76/ 0.67/ 33.0</td>
<td>0.58/ 0.69/ 0.52/ 0.73/ 0.61/ 0.77/ 0.74/ 0.48/ 0.74/ 0.48/ 0.74/ 0.48/ 0.74/ 0.48/ 0.74/ 0.48/ 0.74/ 0.48/ 25.2</td>
</tr>
<tr>
<td>QAMR</td>
<td>0.73/ 0.89/ 0.57/ 0.83/ 0.74/ 0.89/ 0.67/ 0.41/ 0.62/ 0.28/ 0.77/ 0.68/ 0.67/ 0.75/ 0.57/ 0.57/ 58.6</td>
<td>0.62/ 0.73/ 0.61/ 0.85/ 0.79/ 0.89/ 0.75/ 0.37/ 0.56/ 0.31/ 0.72/ 0.67/ 0.56/ 0.78/ 0.76/ 0.67/ 32.7</td>
<td>0.70/ 0.73/ 0.61/ 0.85/ 0.79/ 0.89/ 0.75/ 0.37/ 0.56/ 0.31/ 0.72/ 0.67/ 0.56/ 0.78/ 0.76/ 0.67/ 33.0</td>
<td></td>
</tr>
<tr>
<td>FBC</td>
<td>0.80/ 0.90/ 0.83/ 0.90/ 0.90/ 0.90/ 0.90/ 0.66/ 0.66/ 0.78/ 0.78/ 0.78/ 0.78/ 0.78/ 0.78/ 0.78/ 0.78/ 0.78/ 32.7</td>
<td>0.57/ 0.63/ 0.62/ 0.68/ 0.67/ 0.75/ 0.57/ 0.57/ 58.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CaSt-19</td>
<td>0.51/ 0.23/ 0.70/ 0.61/ 0.57/ 0.73/ 0.68/ 0.61/ 65.0</td>
<td>0.51/ 0.23/ 0.70/ 0.61/ 0.57/ 0.73/ 0.68/ 0.61/ 65.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Transferability test scores using ISEEQ-ERL to answer RQ3. **gray cell:** ISEEQ-ERL trained and tested on same dataset. **dark gray cell:** shows acceptable cross-domain {Train-Test} pairs, where train size is smaller than test size.

**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 6, 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} are generated far better ISQs for IS queries.

12-S-BERT Pairwise similarity between retrieved passages in QAD, QAMR, FBC, and CaSt-19 is <9%, which is minimal.


