KGConv, a Conversational Corpus grounded in Wikidata

Quentin Brabant[1*], **Lina M. Rojas-Barahona**[1], **Gwénolé Lecorvé**[1*], **Claire Gardent**[2]
(1) Orange Innovation
2 Avenue Pierre Marzin, Lannion, France.
{quentin.brabant, gwenole.lecorve, linamaria.rojasbarahona}@orange.com
(2) CNRS/LORIA, Université de Lorraine
Nancy, France.
claire.gardent@loria.fr
(*) equal contribution

Abstract

We present KGConv, a large corpus of 71k English conversations where each question-answer pair is grounded in a Wikidata fact. The conversations were generated automatically: in particular, questions were created using a collection of 10,355 templates; subsequently, the naturalness of conversations was improved by inserting ellipses and coreference into questions, via both handcrafted rules and a generative rewriting model. The dataset thus provides several variants of each question (12 on average), organized into 3 levels of conversationality. We provide baselines for the task of Knowledge-Based Conversational Question Generation. KGConv can further be used for other generation and analysis tasks such as single-turn question generation from Wikidata triples, question rewriting, question answering from conversation or from knowledge graphs and quiz generation.

Keywords: Dialogue, Knowledge Base, Conversational Question Generation

1. Introduction

Unlike open domain and task-oriented dialogues, information seeking conversations are driven by the desire to acquire or evaluate knowledge. These conversations are central for instance, in educational (tutoring a student about a given topic by asking her a set of questions about that topic) and entertainment (quizzes) settings. As large knowledge graphs such as Wikidata[1] have started to emerge, recent years have seen an increasing interest in developing datasets and conversational question answering models that can support such information seeking interactions by grounding conversations in factual data. However, these often focus on question answering (QA) (Saha et al., 2018) or provide datasets of restricted size and variety however (Christmann et al., 2019; Lecorvé et al., 2022).

In this paper, we focus on information seeking conversations where, as illustrated in Table 1, each question-answer turn is grounded in a single fact. Our contribution is two fold. First, we make available the KGConv dataset[2]online where each question-answer pair is grounded in a Wikidata fact. To create a diverse, large scale dataset (70k conversations), we develop conversations for eight different topics (Country, Food, Person, Religion/Ideology, Space Object, Taxon[3], Molecular Entity, and Historical Event). One originality of KGConv is that it provides each question in several versions, each version belonging to one contextuality level: C0, C1 or C2: (i) C0 corresponds to questions whose interpretation is independent of the previous turns, (ii) C1 corresponds to C0 questions into which some entity names are replaced by pronouns or alternative labels using a rule based approach and (iii) C2 corresponds to questions obtained by feeding C1 questions to a T5 model trained for rewriting questions into a more conversational form. Table 1 illustrates the notion of contextuality level with an excerpt of a KGConv conversation.

Our second contribution is to establish some baselines for Knowledge-Based Conversational Question Generation (CQG), the task of generating a question given both a Knowledge graph (KG) fact and a conversational context. While much previous work has focused on Knowledge-Based, conversation Question Answering (Saha et al., 2018; Perez-Beltrachini et al., 2023) or on context-independent, knowledge-based question generation (Bordes et al., 2015; Elsahar et al., 2018; Han et al., 2022), we provide a first investigation of how knowledge-based question generation interacts with conversational context. We report results using both automatic metrics and human evaluation.

2. Related Work

Question generation from RDF triples is adressed in (Bordes et al., 2015; Elsahar et al., 2018; Han et al., 2022) and from small KGs depicting multi-

[1] https://www.wikidata.org/

[2] https://github.com/Orange-OpenSource/KGConv/

[3] A *taxon* is a population, or group of populations of biological organisms, e.g. lions or dinosaurs.

LREC-COLING 2024, pages 9732–9742
20-25 May, 2024.

Triples
Q1: (*NGC 4833*, *part of*, *Milky Way*)
A1: Milky Way
Q2: (*NGC 4833*, *discoverer or inventor*, *Nicolas Louis de Lacaille*)
A2: Nicolas Louis de Lacaille
Q3: (*Nicolas Louis de Lacaille*, *religion or worldview*, *Catholic Church*)
A3: Catholic Church
Contextuality level 0 (C0)
Q1: NGC 4833 is part of what astronomical object?
A1: Milky Way
Q2: What was the name of the discoverer of NGC 4833?
A2: Nicolas Louis de Lacaille
Q3: What was Nicolas Louis de Lacaille's religion?
A3: Catholic Church
Contextuality level 1 (C1)
Q1: NGC 4833 is part of what astronomical object?
A1: Milky Way
Q2: What was the name of the discoverer of NGC 4833?
A2: Nicolas Louis de Lacaille
Q3: What was his religion?
A3: Catholic Church
Contextuality level 2 (C2)
Q1: NGC 4833 is part of what astronomical object?
A1: Milky Way
Q2: Who discovered this object?
A2: Nicolas Louis de Lacaille
Q3: What was his religion?
A3: Catholic Church

Table 1: Excerpt of a conversation at the triple, C0, C1, and C2 levels. The C0-C1-C2 variants of a question Qi are based on the same template. The root entity is NGC 4833, from the theme "space object".

hop questions (Serban et al., 2016; Kumar et al., 2019; Bi et al., 2020) and recently in LC-QuAD 2.0 (Dubey et al., 2019) and ParaQA (Kacupaj et al., 2021). However, these works are limited to the generation of isolated questions, thus no conversational context is under consideration.

CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018) and Wizard-of-Wikipedia (Dinan et al., 2019) are conversational QA corpora, in which answers are extracted from paragraphs, instead of KGs. Similar to KGConv, ConvQuestions (Christmann et al., 2019) and CSQA (Saha et al., 2018) are conversational corpora based-on structured knowledge. However, the former does not provide the triples for the questions and contains only 315 distinct conversations. The latter, despite being a very large dataset covering a wide range of questions types[4], contains rather unnatural questions with a proprietary formalism which does not directly correspond to Wikidata triples. Recent work proposes to generate questions from SPARQL queries, especially to express complex questions (Lecorvé et al., 2022; Perez-Beltrachini et al., 2023). However, manually annotating questions with SPARQL queries is difficult and the authors turn on a small set of 350 reference questions. In contrast, our dataset focuses on triples and simple questions which enables a large and varied set of reference questions, obtained through manually written verbalization templates.

Although OpenDialKG (Moon et al., 2019) also provides conversations grounded in KG, conversations in this corpus are free-form. Additionally, mentions are grounded by entities while conversation transitions are grounded by properties in the KGs. In contrast, KGConv, contains tutoring conversations wherein every question-answer pair is grounded by a fact represented by a KG triple. A key limitation of OpenDialKG however is that dialogues are linked to FreeBase, a Knowledge Base which is no longer publicly available.

Finally, our work capitalizes on two existing corpora which provide a correspondence between triples and questions, namely SimpleQuestions (Bordes, Antoine and Usunier, Nicolas and Chopra, Sumit and Weston, Jason) and ZeroShot Relation Extraction (Levy et al., 2017). In comparison, KGConv extends the question templates from these corpora by proposing 3,879 new templates, and focusing on 458 properties. Furthermore, while these two corpora focus on isolated questions or short follow-up turns of maximum three turns, KGConv contains 8.7 turns per conversation on average.

3. Overview of the Dataset

In KGConv, each conversation is focused on a given *root* entity. As illustrated in Table 1, the first question bears directly on this root entity, while further questions explore new facts about any entity discovered during the conversation (including the root entity itself). Hence, a conversation can be seen as a small evolving KG, where each turn expands the conversation graph with a new entity and the property which connects it to the existing graph.

For each root entity, three conversations are derived from Wikidata in order to increase the diversity of the dataset. The corpus covers eight themes: Country, Food, Person, Religion/Ideology, Space Object, Taxon[5], Molecular Entity, and Historical Event. The theme of a conversation corresponds to

[4]Single or multiple triples, entity/numeric/boolean answers, comparative questions

[5]A *taxon* is a population, or group of populations of biological organisms, e.g. lions or dinosaurs.

the Wikidata class associated to the root entity e.g., Person corresponds to the Q215627 class in Wikidata. Table 2 summarizes the size of the dataset for each theme. We use Taxon and Space Object as unseen themes, which means they are not seen at training time. Our test data also include a set wherein each conversation contains at least one of 85 "unseen" properties that do not appear in the train and dev data. These properties are referred as "with unseen prop" in Table 2. The number of questions in a conversation is at least 5, at most 19 and 8.6 on average. In total, KGConv has $70,596$ conversations including $603,905$ questions about $63,345$ distinct Wikidata entities and 458 properties. To enable links with Wikidata and further extensions, Wikidata IDs are provided for all entities and properties along with their natural language labels. In addition, each question in a conversation has 6.8 paraphrases on average, and each paraphrase has three versions that correspond to three levels of contextualization:

- C0: this version of a question is produced automatically from a KG fact independent of the conversation context.
- C1: this version is derived from the C0 version taking the context into account and using rules to substitute repetitions with anaphoric forms.
- C2: this version is derived from the C1 version by applying a generative model trained to rewrite questions.

4. Data Collection

The process for creating the corpus is summarized in Figure 1 and elicited in the following subsections.

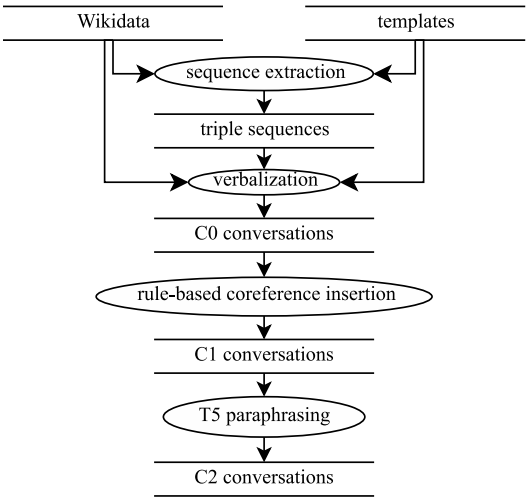

Figure 1: Data flow diagram of the dataset creation.

4.1. Sequence Extraction

In this step, we extract sequences of triples from Wikidata which will be used to ground the conversations.

We first extract triples in which:

1. the subject has at least one English label and the object either has an English label or is a literal (string, number, date, etc.);
2. the property belongs to a pre-defined set of Wikidata properties which excludes "uninteresting" properties. In particular, we excluded properties whose prefix is not "`wdt:`" (to avoid triples that provide meta-information), and properties which link entities to their ID in some other database.

For any triple (s, p, o) in this set, we then create the reversed triple $(o, -p, s)$ where $-p$ denotes the inverse of property p. In this way, if our subgraph contains (*France*, *capital*, *Paris*), it also contains (*Paris*, *capital*, *France*) which permits creating questions about both the subject s and the object o e.g., "What is the capital of France" and "Paris is the capital of which country?". We call the resulting Wikidata subgraph W for World.

Based on this set of triples, we then create sequences of triples as follows. Each conversation will focus on triples in the neighborhood $\mathcal{N}(r)$ of a root entity r in W. This neighborhood is defined as the subgraph of W containing r and all nodes (i.e. entities) that are 1 or 2 edges (i.e. properties) away from r; in other words, $\mathcal{N}(r)$ contains all triples of the form (r, p, o_1) and their successors of the form (o_1, q, o_2). Roots were sampled from instances of the 8 themes displayed in Table 2, with the condition that their neighborhood is large enough (at least 20 triples) to generate 3 reasonably long conversations with enough differences.

For each root, 3 triple sequences of the form $(t_0, t_1, \ldots, t_n)$ were built iteratively by picking triples from $\mathcal{N}(r)$ in a greedy stochastic process. At each step of the process, the subject of the chosen triple is either the root (i.e. $s_i = r$) or an entity, either subject or object, from the previous triple ($s_i = s_{i-1}$ or $s_i = o_{i-1}$). Additionally, a triple cannot appear twice in the sequence. The decision to stop or continue the process is made at each time step i following a probability that increases with i: $\mathsf{Pr}_{stop}(i) = 0.06i - 0.18$.

4.2. Verbalization

Questions are generated using templates like "What is the capital of ___?", where the slot is to be filled by the subject of a triple. Each template τ is applicable for a given property p_τ (e.g.,"capital of"), given the required types $\mathcal{S}_\tau$ for the subject slot and

	entities	prop.	triples	conv.	number of question-turns				templates	references
	(number in all conversations)				train	dev	test	total	(avg. per question-turn)	
person	32k	327	72k	26k	185k	29k	11k	226k	7.2	12.5
country	2k	171	3k	0.7k	5k	0.8k	0.2k	6k	5.5	9
ideology	1k	169	2k	0.4k	3k	0.6k	0.2k	4k	6.6	11.4
space object	3k	116	6k	6k	0	0	50k	50k	7.3	12.4
molecular entity	18k	151	38k	23k	155k	25k	10k	189k	6.4	11.7
historical event	5k	189	8k	5k	35k	6k	2k	43k	5.6	10.2
food	3k	166	4k	2k	15k	2k	1k	18k	6.1	10.4
taxon	3k	215	5k	2k	0	0	16k	16k	7.9	13.6
with unseen prop.	14k	404	24k	6k	0	0	52k	52k	6.9	12.1
whole dataset	63k	458	143k	71k	398k	63k	143k	604k	6.8	12

Table 2: Quantitative summary of KGConv. Entities, properties and triples can appear in several conversations and several themes but are only counted once. The two last columns show the average number of templates used in a single question-turn, and the number of distinct references (including C0, C1 and C2 versions of all template-based verbalizations).

the required types $\mathcal{O}_\tau$ for the object, which will be the answer. This applicability condition on τ is denoted as $C(\tau) = (p_\tau, \mathcal{S}_\tau, \mathcal{O}_\tau)$. Then, a triple (s, p, o) satisfies $C(\tau)$ if: $p_\tau = p$, $\mathcal{S}_\tau \subset \text{types}(s)$, and $\mathcal{O}_\tau \subset \text{types}(o)$.

To create a large number of diverse questions for all properties in W, we gathered templates from three sources. Table 3 summarizes the number of templates and their sources. The following sections provide details on the methods used to get templates from each source.

4.2.1. Zero-Shot templates

The Zero-Shot dataset (Levy et al., 2017) contains 1,192 question templates spanning 120 Wikidata properties. Each template τ was originally created to ask a question for a given property p, regardless the type of the subject and the object. Templates grounded on properties that were no longer in Wikidata were discarded. For the remained templates, we defined the applicability condition as $C(\tau) = (p, \emptyset, \emptyset)$.

4.2.2. Simple questions v2 templates

We automatically extracted templates from the SimpleQuestions_v2 dataset (Bordes et al., 2015), which contains 108k triple-question pairs, involving 131k distinct entities and 1,837 properties.

As SimpleQuestions_v2 is based on the Freebase KG, we translated entities and properties into their Wikidata counterpart: we relied both on the Wikidata property P646, that links Wikidata entities to their Freebase counterpart, and on an available mapping between Freebase and Wikidata properties[6]. This allowed us to get Wikidata counterparts for 83,447 entities and 142 properties.

[6] https://www.wikidata.org/wiki/Wikidata:WikiProject_Freebase/Mapping

We then extracted templates from question-triple pairs whose triple could be translated to Wikidata. For each such triple-question pair $((s, p, o), q)$, we created a template τ by replacing in q the label of s by an empty slot. The applicability condition of τ was then defined as $C(\tau) = (p, \text{types}(s), \text{types}(o))$, where $\text{types}(x)$ is the set of all types of x in Wikidata. Since only a small subset of Freebase entity and triples could be translated into a Wikidata counterpart, many triples were filtered out, so the extracted templates only cover 77 of our properties.

4.2.3. New templates

We manually created additional templates in three steps: (1) extracting applicability conditions, (2) writing templates corresponding to these conditions, (3) validating written templates.

Step 1: Extracting applicability conditions. From the neighbourhoods of potential roots of all themes, we gathered a set $\{(s_i, p_i, o_i)\}_{i=0}^N$ of Wikidata triples for which we had no template corresponding to p_i. From this set of triples we generated a set of applicability conditions $\{(p_i, \text{types}(s_i), \text{types}(o_i))\}_{i=0}^N$ to be annotated with corresponding templates. This resulted in many applicability conditions, with overly specific subject and object types for each property. We solved this problem by iteratively merging conditions with the same property, via semi-automated conceptual agglomerative clustering process. Merging two conditions $(p, \mathcal{S}_i, \mathcal{O}_i)$ and $(p, \mathcal{S}_j, \mathcal{O}_j)$ consists in replacing them with a new one $(p, \mathcal{S}_i \cap \mathcal{S}_j, \mathcal{O}_i \cap \mathcal{O}_j)$, which necessarily is matched by more triples than the two original ones. At the end of this process, applicability conditions that were met by less than 5 triples were discarded.

	Properties	Temp.	Temp. per prop. min / avg /max.
SimpleQuestions	77	5,817	1 / 76 / 453
ZeroShot RE	75	771	1 / 10 / 31
New templates	413	3,879	1 / 9 / 80
Total	474	10,355	1 / 22 / 453

Table 3: Statistics on properties and templates for each sources of templates.

Step 2: Template writing. The next step was to write, for each applicability condition C, templates that would apply to any triple matching C.

Three students of an NLP Master program were hired to annotate the question templates. They were native English speakers hired on a short term contract to perform various annotation tasks for NLP. They were paid slightly above the national minimum wage. We provided them with an annotation tool, in which one applicability condition at a time was displayed, along with 5 examples of matching triples. While annotators were writing templates, their results on these triple were displayed. To speed up this process, artificially generated templates were also proposed to the annotators, who could accept or reject them. Afterward, accepted artificial templates were treated in the same way as those written by a human annotator.

Step 3: Template validation. To ensure the quality of the resulting templates, all of them were manually filtered by the authors.

4.3. Contextualization

Applying the templates from Section 4.2 to the triple sequences from Section 4.1 yields conversations where questions are not contextualized (C0). To improve the naturalness of the conversations, we derived two in-context versions from these C0 conversations. The first one, (C1), is obtained by applying hand-crafted rules to introduce coreferences and correct some errors produced during verbalization with templates. The second version (C2), results from rewriting C1 questions with a T5 model trained to rewrite questions.

4.3.1. Post-Processing Conversations (C1 Variants)

Referring Expressions. In Wikidata, an entity can have several labels: one of those is called "preferred label" and is meant to be used by default. In the C0 version, entities are always referred by their preferred label. This step introduces variability by replacing some of the preferred labels with other available labels from Wikidata, according to the following rules: (1) the first reference to an entity in the conversation is the preferred label, or contains it as a substring; (2) further mentions are labels that are substrings of the first reference. For instance, the entity Q9592 has the preferred label $l_1 =$ "Catholic Church", alternative labels $l_2 =$ "Roman Catholic Church" and $l_3 =$ "Roman Apostolic Catholic Church"; if l_2 is used as the first reference to the entity, next references will use either l_1 or l_2 but not l_3, since it is not a substring of l_2. Whenever the subject is a person with a name and surname, we include its surname in the set of available labels.

Determiners. Deciding which label should be preceded by "the" is not trivial. For example, "United Kingdom" and "Republic of China" require it, while "France" and "China" do not. This step handles this problem by asking a BERT language model[7] to fill a mask token inserted before the label; when "the" was predicted with a probability at least 0.92, it was inserted before the label.

Tense. We noticed that most templates are written in present tense, while many triples describe facts that are no longer true or concern dead people, past events, etc. Questions were rewritten in the past tense[8] if the corresponding triple had an "end time" qualifier in Wikidata, or if its subject or object was a dead person.

Rule-based introduction of pronouns. Subject mentions are pronominalised using a rule-based approach: a pronoun is used only if the subject also appears in the triple of the previous question and if its gender differs from the gender of the object of this triple (to avoid ambiguous pronouns); further rules are used to determine the kind of pronoun to use (for example, if the subject reference is followed by a possessive "s", a possessive pronoun should be used, etc.).

4.3.2. Model-based rewritings (C2 Variants)

To further increase the contextuality of questions, a T5-based question rewriting model[9] was fine-tuned on a training set derived from 2 conversational machine reading QA datasets, namely CANARD (Elgohary et al., 2019) and CoQAR (Brabant et al., 2022). This training set is made of 142K instances. For each instance, the input is a question q_i, along with its conversation history $[q_0, a_0, \ldots, q_{i-1}, a_{i-1}]$,

[7] https://huggingface.co/bert-base-uncased

[8] https://github.com/bendichter/tenseflow

[9] From the base version on HuggingFace: https://huggingface.co/t5-base.

while the output is a semantically equivalent question whose form is expected to be natural in a conversation, denoted by q_i^*. In some instances, q_i and q_i^* have respectively a C0 and a contextualized form. In other instances, q_i and q_i^* are equal; these instances correspond to cases where either q_i already has an C1 form, or there is no natural way to rewrite it without losing information or bringing ambiguity. Including such cases to the training set enables the model to learn when it should rewrite the input question or not.

At inference time, the 20 best hypotheses are generated by the model for each instance. Then, they are classified into three authorized categories, using a set of expert conditions: (1) coreference with a pronoun (e.g., "In which country is Kyoto located?" rewritten as "In which country is *it* located?"), (2) coreferences with a demonstrative noun phrase (e.g., "In which country is *this city* located?"), and (3) ellipses (e.g., "In which country?"). Those that do not belong to any category are filtered out; moreover, to limit possible ambiguities, we prohibit two consecutive reformulations of the same category. Finally, if some hypothesis remain, the one with the highest probability is selected as the rewritten form.

This process was applied on all C1 questions, leading to the C2 version shown in Table 4.

5. Conversational Question Generation

Knowledge-Based, Conversational Question Generation extends Question Generation from Knowledge Graph triples (Elsahar et al., 2018; Han et al., 2022) to a conversational setting: instead of generating a question only from a triple, we generate a question from both a triple and the preceding conversational context. This raises the additional challenge of generating questions in contextually appropriate forms e.g., using appropriate referring expressions and ellipses. Leveraging the multimodal text/graph nature of our dataset, we explore four ways of representing the context: (1) no contextual information at all (Empty), (2) the sequence of previous questions and answers (NL) (3) the sequence of triples underlying the questions and answers (KG) and (4) the sequence of questions and answers with their corresponding triples (NL+KG).

For each of the four variants, we trained a baseline by fine-tuning a T5-small model on the three versions of questions in KGConv (C0, C1 and C2). In the train and dev sets, all themes are mixed together. The number of epochs for training is determined via early stopping.

Type	Original	Rewritten	OK?
Coref. w/ pronoun	Which location is Switzerland a component of?	Which location is it a component of?	✓
	What was the cause of death of Uriella?	What was her cause of death?	✓
	What title was held by Martin of Tours?	What title was held by him?	✓
	Who is in charge of the government of Warsaw?	Who is in charge of the government there?	✓
	Pierre Chaunel was who's spouse?	Was Chaunel his spouse?	×
Coref. w/ demonstrative noun phrase	With which country would you associate Gyeonggi Province?	With which country would you associate this province?	✓
	Which reference work outlined Albigensian Crusade?	Which reference work outlined this conflict?	✓
	Where are World Council of Churches's headquarters?	Where are these headquarters?	×
Ellipsis	What is the public holiday associated with Switzerland?	What is the public holiday?	✓
	What is the zenith of Eritrea?	What is the zenith?	✓
	In what geographic region is Eurasia located?	In what geographic region?	✓

Table 4: Examples of rewritten questions at inference time.

Root entity type	C1 OOC	C1 Shortened label	C1 Coref. w/ poss. pron.	C1 OOC	C1 Shortened label	C1 Coref. w/ poss. pron.	C1 Coref. w/ dem. pron.	C1 Ellipsis
person	62.4	13.8	23.7	26.6	8.8	44.8	4.6	15.2
country	87.5	4.1	8.4	43.3	2.0	23.3	9.4	22.0
ideology	81.3	6.4	12.3	28.6	3.2	37.3	9.1	21.8
space object	84.6	1.6	13.7	25.8	1.0	40.5	18.8	13.9
molecular entity	83.2	10.8	6.0	22.5	8.2	33.7	19.5	16.1
historical event	84.9	6.7	8.4	23.2	2.9	35.6	13.7	24.5
food	82.9	7.6	9.4	26.9	3.8	36.6	8.1	24.5
taxon	83.8	2.0	14.2	28.1	1.1	35.7	16.0	19.2
w/ unseen prop.	67.1	12.6	20.2	27.0	8.2	41.4	7.8	15.7

Table 5: Percentage of questions using an alternative label and a pronoun in C1 questions; percentage of C2 questions that differ from the C1 version.

6. Automatic Evaluation

We evaluate each model on the test set using Google-BLEU and BERT-score taking as references all questions associated with the input triples (in C0, C1, and C2 forms), around 12 references

on average. The results are presented in Table 6.

Seen vs. unseen. Unsurprisingly, all models obtain lower scores on unseen themes. Similarly, average scores are lower on unseen properties because the verbalization highly depends on the property of the triple.

C0 vs. C1 vs. C2. In term of GLEU, models trained on C2 generally perform better than their C1 and C0 counterparts. However, this tendency is not confirmed with BERT-scores. Moreover, these better GLEU scores might just be an artifact of the experimental design. Indeed references contain C0, C1 and C2 question versions, and models trained on C2 are the only one to have been trained to generate all three versions. Thus, C2 models might have better GLEU simply because their training data is more in-line with the references.

Conversational context format. Adding conversational context to the models trained on C1 and C2 questions consistently improves GLEU and BERT-scores. Looking at GLEU scores, it also seems that providing the context in the form of triples (or triples and text) provides a better improvement than providing the context in the form of text.Since conversational context is not required to generate C0 questions, models trained on C0 questions tend to perform better when no conversational context is given, except for unseen properties.

7. Human Evaluation

The human evaluation assesses both the dataset and the two baselines: C1 (KG+NL) and C2 (KG+NL). We sample conversations from the test set. For each conversation we created four alternative versions. The first two versions select C1 and C2 questions from the dataset (i.e., the references). The other two versions turn on the questions C1 and C2 generated by the baselines that takes into consideration both the triple and the context. We built an evaluation graphical interface, in which human evaluators can rate these conversations. Table 7 gives the number of conversations rated by evaluators depending on theme and version.

7.1. Evaluation Setup

The ratings were provided by 15 evaluators from the authors' research center (excluding the authors). Each evaluator could evaluate up to 50 conversations. They were told that conversations are automatically generated, but were provided no information about the method employed. Conversations were presented one by one to the evaluator. For each question-answer pair, the corresponding triple was displayed, and the evaluator had to (i) rate the linguistic correctness of the question on a 5 point scale and (ii) evaluate whether the question-answer pair expresses the information of the triple ("yes", "quite", "no", "I don't know"). In addition, the evaluator had to rate the naturalness of the whole conversation flow.

A second round of evaluation on 61 of the already rated conversations was performed to assess the consistency of ratings among annotators. This evaluation utilizes the C1 and C2 versions of the questions (i.e., conversations using baselines were not included in the second round). This evaluation was performed by 3 annotators who did not participate in the first round.

7.2. Results

Results are provided in Table 8. We use the Mann-Whitney U and the χ^2 test to assess significance. The former was used for correctness, clearness and naturalness scores, since those are evaluated on an ordered scale. The latter was used for faithfulness scores, since these form a scale that is not completely ordered (because of the "I don't know" answer).

C1 vs C2 references. Comparing the scores of C1 and C2 from the references (in the *All* block of Table 8) we see that, while linguistic correctness is roughly the same, C2 references are less clear (clearness 4.59 vs 3.96, p=2e-12), less faithful to the triples (0.90 yes vs 0.71, p=8e-8) and less natural overall (3.88 vs 3.30, p=0.006).

C1 vs C2 baselines. The C2 baseline seems a bit more linguistically correct (4.60 vs 4.75, p=0.001). However, the C1 baseline seems clearer (4.56 vs 4.13, p=4e-5) and more natural overall, although it might be due to chance (3.64 vs 3.14, p=0.066). Faithfulness seems to be the same.

References vs baselines. Now let us compare the baselines to the references they were trained on. For the C1 reference, the baseline is less faithful to triples (0.90 yes vs 0.78, p=0.00014), otherwise we observe no significant difference. For the C2 reference, the baseline is better on every measure, although we obtain low p-values only for clearness (3.96 vs 4.13, p=0.012) and correctness (4.59 vs 4.75, p=0.0012).

Seen vs. unseen. The scores obtained on unseen themes tend to be lower that those obtained on seen themes. This difference happens both for the C1 baseline and for the C2 reference. This suggests that the differences are due to the difficulty

	Training data:	C0				C1				C2			
	Conversational context:	Empty	KG	NL	KG+NL	Empty	KG	NL	KG+NL	Empty	KG	NL	KG+NL
GLEU	Seen themes	**0.482**	0.476	0.480	0.467	0.464	**0.491**	0.472	0.485	0.488	0.519	0.522	**0.535**
	Unseen themes	**0.409**	0.390	0.396	0.365	0.395	0.411	0.406	**0.417**	0.396	**0.457**	0.429	0.442
	Unseen prop.	0.194	**0.214**	0.209	0.200	0.192	0.225	0.215	**0.236**	0.210	0.231*	0.226	**0.244**
BERT score	Seen themes	**0.815**	**0.815**	**0.815**	**0.815**	0.816	**0.818**	0.816	0.816	0.815*	0.816*	**0.818**	0.816*
	Unseen themes	**0.793**	0.792	0.792	0.790	0.787	0.794	0.792	**0.796**	0.787	**0.796**	0.787	0.788
	Unseen prop.	0.770	0.773	**0.774**	0.772	0.770	0.777	0.774	**0.779**	0.768	**0.776**	0.775*	0.774

Table 6: **Results of the automatic evaluation.** Seen themes are those with a non-empty training set (see Table 2), unseen themes are space object and taxon. The scores are obtained by macro-averaging over themes. The best score is in bold; lower scores that do not differ significantly ($p > 0.05$ in a Mann-Whitney U test) from the best one are adorned with (*).

		Ref. C1	Ref. C2	Model C1	Model C2
(seen)	person	10	10	9	8
	food	10	11	10	10
(unseen)	taxon	8	8	8	9
	space o.	6	8	9	8

Table 7: Number of rated conversations.

of themes rather that the fact that they were seen during the training of baselines or not.

Inter-rater agreement. We computed Cohen's kappa for each metric (faithfulness, correctness, clearness, naturalness) and obtained, respectively: 0.23, 0.10, 0.22, and 0.14

Although those scores are quite poor, the confusion matrices (Table 9) suggest that, although the exact rate given to a question has a high degree of subjectivity, raters tend to give close ratings.

Low kappas seem due to two factors: (1) the intrinsically subjective nature of the task, which can explain that raters disagree by giving different but close rates, (2) genuine mistakes made by raters (for example, when faithfulness is rated at yes and no by two different raters). It is also possible that differences in raters' fluency had an impact on agreement. Despite the low agreement, we observed the interesting regularities reported previously in this section.

8. Conclusion

We make available KGConv, a new conversational dataset grounded in Wikidata where each question-turn in the conversation comes into several variants belonging to 3 contextuality levels (C0, C1, C2). Although C2 questions have more diverse forms than C1 questions, the results of human evaluation suggest that C1 questions are more reliable than C2 questions.

We also presented several baselines for the task of question generation and found that generating questions from unseen properties is challenging for these baselines. An interesting perspective would be to investigate methods for tackling this particular zero-shot task.

As it provides a large number of references for each question in a conversation, KGConv is well suited for other tasks besides Conversational Question Generation such as in particular, single-turn question generation from facts, question rewriting and generation of sequence of question-answer pairs from a Knowledge graph (KG) or vice-versa.

9. Limitations

This corpus has been generated semi-automatically, although human annotations were involved in the question templates, the conversations were generated automatically from the KG. As a consequence, in some cases the flow of the conversation may be unnatural, because humans do not usually talk in that way. This might be specially true when conversations involve complex content (e.g. molecular entities, space objects or historical events) that may be difficult to be understood by non experts.

			# of conv.	# of quest.	Faithfulness to the triple % yes↑	% no↓	% quite↓	% idk	Linguistic correctness↑	Semantic clearness↑	Conversat. naturalness↑
Seen	Ref.	C1	20	184	**0.91**	**0.05**	**0.03**	0.01	4.61	4.60	**3.85**
		C2	21	183	0.77	0.15	0.08	0.01	4.64	4.09	3.52
	Baselines	C1	19	178	0.84	0.10	0.04	0.02	4.67	**4.65**	3.63
		C2	18	161	0.80	0.12	0.07	0.01	**4.73**	4.21	3.39
Unseen	Ref.	C1	14	128	**0.88**	**0.06**	**0.02**	0.03	4.59	**4.57**	**3.93**
		C2	16	140	0.64	0.19	0.11	0.06	4.51	3.79	3.00
	Baselines	C1	17	157	0.71	0.14	0.13	0.03	4.53	4.47	3.65
		C2	17	153	0.76	0.10	0.12	0.02	**4.76**	4.05	2.88
All	Ref.	C1	34	312	**0.90**	**0.06**	**0.03**	0.02	4.60	**4.59**	**3.88**
		C2	37	323	0.71	0.17	0.09	0.03	4.59	3.96	3.30
	Baselines	C1	36	335	0.78	0.12	0.08	0.02	4.60	4.56	3.64
		C2	35	314	0.78	0.11	0.10	0.02	**4.75**	4.13	3.14

Table 8: Scores from human evaluation for conversations about seen, unseen or all themes. Ref., stands for the C1 and C2 references.

	1	2	3	4	5
1	3	1	3	2	2
2		1	5	5	15
3			3	8	34
4				13	118
5					344

Correctness

	1	2	3	4	5
1	5	9	6	6	10
2		7	11	12	22
3			8	16	43
4				13	78
5					311

Clearness

	1	2	3	4	5
1	0	1	1	0	0
2		3	3	7	1
3			5	10	5
4				14	9
5					2

Naturalness

	yes	quite	no	idk
yes	361	97	33	10
quite		22	16	4
no			13	1
idk				0

Faithfulness

Table 9: Confusion matrices of human ratings.

10. Ethics Statement

The construction of this corpus involved manual annotation of the question templates. Therefore, we hire three students of an NLP Master program. They were native English speakers hired on a short term contract to perform various annotation tasks for NLP. They were paid slightly above the national minimum wage and they had the right to the social security benefits.

11. Bibliographical References

Sheng Bi, Xiya Cheng, Yuan-Fang Li, Yongzhen Wang, and Guilin Qi. 2020. Knowledge-enriched, type-constrained and grammar-guided question generation over knowledge bases. In *Proceedings of the International Conference on Computational Linguistics (CICLing)*, pages 2776–2786.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. ArXiv:1506.02075 [cs].

Quentin Brabant, Gwénolé Lecorvé, and Lina M. Rojas Barahona. 2022. CoQAR: Question rewriting on CoQA. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 119–126, Marseille, France. European Language Resources Association.

E. Choi, H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 729–738. Association for Computing Machinery.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *Proceedings of the The Semantic Web (ISWC)*, pages 69–78. Springer International Publishing.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. 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)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Hady Elsahar, Christophe Gravier, and Frederique Laforest. 2018. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 218–228, New Orleans, Louisiana. Association for Computational Linguistics.

Juliette Faille, Albert Gatt, and Claire Gardent. 2021. Entity-based semantic adequacy for data-to-text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1530–1540, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelvin Han, Thiago Castro Ferreira, and Claire Gardent. 2022. Generating questions from Wikidata triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 277–290, Marseille, France. European Language Resources Association.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133, Prague, Czech Republic. Association for Computational Linguistics.

Endri Kacupaj, Barshana Banerjee, Kuldeep Singh, and Jens Lehmann. 2021. ParaQA: A Question Answering Dataset with Paraphrase Responses for Single-Turn Conversation. In *Proceedings of the The Semantic Web (ISWC)*, pages 598–613. Springer International Publishing.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 97–102.

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-Controllable Multi-hop Question Generation from Knowledge Graphs. In *Proceedings of The Semantic Web (ISWC)*, pages 382–398. Springer International Publishing.

Gwénolé Lecorvé, Morgan Veyret, Quentin Brabant, and Lina M. Rojas-Barahona. 2022. SPARQL-to-text question generation for knowledge-based conversational applications.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679.

Laura Perez-Beltrachini, Parag Jain, Emilio Monti, and Mirella Lapata. 2023. Semantic parsing for conversational question answering over knowledge graphs. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2507–2522, Dubrovnik, Croatia. Association for Computational Linguistics.

S. Reddy, D. Chen, and C.D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Iulian Vlad Serban, Alberto García-Durán, Çaglar Gülçehre, Sungjin Ahn, Sarath Chandar, Aaron C Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL)*.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

12. Language Resource References

Bordes, Antoine and Usunier, Nicolas and Chopra, Sumit and Weston, Jason. *Simple Questions (v2)*. Facebook. PID https://huggingface.co/-datasets/simple_questions_v2.

Levy, Omer and Seo, Minjoon and Choi, Eunsol and Zettlemoyer, Luke. 2017. *Zero-Shot Dataset*. University of Washington Natural Language Processing (UWNLP). PID http://nlp.cs.washington.edu/zeroshot/.