Knowledge Graph Completion-based Question Selection for Acquiring Domain Knowledge through Dialogues

Kazunori Komatani
Yuma Fujioka
Keisuke Nakashima
komatani@sanken.osaka-u.ac.jp
fujioka@ei.sanken.osaka-u.ac.jp
nakashima@ei.sanken.osaka-u.ac.jp
The Institute of Scientific and
Industrial Research (SANKEN)
Osaka University
Ibaraki, Osaka, Japan

Katsuhiko Hayashi
khayashi0201@gmail.com
Gunma University
Maebashi, Gunma, Japan

Mikio Nakano*
mikio.nakano@c4a.jp
Honda Research Institute Japan Co.,
Ltd.
Wako, Saitama, Japan

ABSTRACT

Building a perfect knowledge base in a certain domain is practically impossible, so it is effective for dialogue systems to acquire knowledge for enhancing an imperfect knowledge base through natural language dialogues with users. This paper proposes a framework for selecting questions for such knowledge acquisition when a knowledge graph is used as the knowledge base. The framework uses knowledge graph completion (KGC) for predicting new links that are likely to be correct and selects questions on the basis of the KGC scores. One of the problems with this framework is that questions with incorrect content might be selected, which often occurs when the link prediction performance is low, and this would reduce the users' willingness to engage in dialogues. To alleviate this problem, this paper presents two modifications to the KGC training: 1) creating pseudo entities having substrings of the names of the entities in the graph so that the entities whose names share substrings are connected and 2) limiting the range of negative sampling. Cross validation-based experiments we conducted showed that these modifications improved KGC performance. We also conducted a user study with crowdsourcing to investigate the subjective perception of the correctness of the predicted links. The results suggest that the model trained with the modifications is capable of avoiding questions with incorrect content.

CCS CONCEPTS

• **Computing methodologies** → *Discourse, dialogue and pragmatics*.

*Currently with C4A Research Institute, Inc.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IUI '21, April 14–17, 2021, College Station, TX, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8017-1/21/04...\$15.00

<https://doi.org/10.1145/3397481.3450653>

KEYWORDS

dialogue systems, domain knowledge acquisition, knowledge graph completion

ACM Reference Format:

Kazunori Komatani, Yuma Fujioka, Keisuke Nakashima, Katsuhiko Hayashi, and Mikio Nakano. 2021. Knowledge Graph Completion-based Question Selection for Acquiring Domain Knowledge through Dialogues. In *26th International Conference on Intelligent User Interfaces (IUI '21)*, April 14–17, 2021, College Station, TX, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3397481.3450653>

1 INTRODUCTION

Many information service systems use knowledge bases in their service domains. For example, a movie recommendation system uses a movie database and a restaurant search system uses a restaurant database. Building such a knowledge base requires much effort, so a collective intelligence approach is often taken. With the help of many people, such systems can acquire knowledge and enhance their knowledge base.

The spread of mobile phones and smart speakers has increased the use of UIs that have natural language dialogue functions [21, 40, 45]. Some people prefer a dialogue interface capable of natural language to graphical user interface such as a Web form, because it can be easily used in a mobile environment. We consider that using natural language dialogue is a promising approach for the system to acquire new knowledge. One of the advantages of natural language dialogue is that people can enjoy conversations with the system while telling it new information.

In this study, we assume that a dialogue system has a knowledge graph as the backend database. A knowledge graph is composed of a set of triplets, each of which comprises two entities and a relation, and represents relationships among entities [2]. The triplet is written in the form (*subject, relation, object*). For example, the fact that *Tom yum goong* tastes spicy is written as (Tom yum goong, tastes, spicy). The system engages in dialogues by using the knowledge in the knowledge graph and also tries to store knowledge acquired during dialogues in the graph.

To acquire knowledge, we propose a framework that uses knowledge graph completion (KGC) [6, 27] for predicting new links (triplets) from the embedding of the existing knowledge graph.

The dialogue system asks the user if the predicted link is correct, and the system adds it to the knowledge graph if the answer is correct. There are two sub-problems in this framework. One is to determine which link candidates the system should ask the user whether they are correct, and the other is to determine when to ask about each link candidate depending on the context of dialogues. In this study, we will focus only on the former and leave the latter for future work.

One issue we need to consider is that if the link candidate is incorrect, it may adversely affect the user's impression as shown in [23] and decrease the user's willingness to engage in the dialogue. To avoid this problem, we use the scores of a KGC such as ComplEx [38] for each predicted link that indicate how likely it is the predicted link to be correct. By not asking about the predicted links whose scores are low, the system can avoid incorrect questions.

In this framework, it is assumed that the KGC scores are reliable. However, we found that, when we used a new knowledge graph that was built from scratch, the scores were not good enough to avoid incorrect questions. Therefore, we should improve the reliability of the KGC scores before the questions are made. This paper presents two modifications to the training the KGC model that improve the reliability of the scores. One is to generate pseudo entities representing substrings of the labels of the existing entities. This makes it possible to connect entities whose labels share substrings and results in higher KGC accuracy. The other is to put limitations on negative sampling in ComplEx [38], a KGC technique that we use.

We evaluated the two modifications using a knowledge graph manually built for a chatbot in the food and restaurant domain. We evaluated the KGC results from objective measurements and found that the model trained with the modifications outperformed a baseline without the modifications. We also examined the kind of questions that are obtained when we used the whole knowledge graph as the training data of KGC and found that our model with the modifications could select reasonable questions in terms of correctness; i.e., it had better user impressions.

The contributions of this paper are summarized as follows:

- (1) We propose a framework of using KGC scores to select questions for knowledge acquisition through natural language dialogues.
- (2) We present two modifications to improve the reliability of KGC scores and to avoid questions that may decrease the user's willingness to engage in the dialogue and show their effectiveness through objective evaluations.
- (3) We show example questions that were obtained by applying KGC to a knowledge graph and were selected by its scores. We also show that many reasonable questions about originally unknown triplets were selected through a subjective evaluation.

2 RELATED WORK

2.1 Dialogue Systems Using Knowledge Graphs

Knowledge graphs have recently gained much attention in computational linguistics and are also used in dialogue systems. They are used for understanding user utterances [8, 10, 19], generating better system responses in neural dialogue response generation

[41, 42, 44, 46], and making recommendations in conversational systems [9, 47]. These studies use fixed knowledge graphs. To extend the scope of these kinds of studies, we explored the idea of enhancing an imperfect knowledge graph through dialogue. A recent study used KGC to generate questions relevant to the current topics [43]. However, the authors of that study did not discuss the correctness of the questions.

2.2 Knowledge Acquisition in Dialogue Systems

There is great interest in developing means for computers to learn knowledge autonomously. A well-known example is the Never-Ending Language Learner (NELL) [7, 31], which continuously extracts information from the Web. Several methods have been developed for machine learning tasks to continuously improve the performance of classifiers in a semi-supervised manner, which is known as life-long learning [11].

Unlike these studies, we aim to develop systems that can acquire knowledge through dialogues. Several studies have been conducted on this topic. Most of them attempted to acquire lexical and ontological knowledge [17, 30, 34, 37]. For example, Otsuka et al. [34] proposed a method of estimating the cuisine of an unknown restaurant from the character sequence of its name and accordingly changing question forms to acquire knowledge. Pappu and Rudnicky [35] designed strategies for asking users questions in a goal-oriented dialogue system and analyzed the acquired knowledge through a user study. Hixon et al. [20] proposed a method for asking questions to obtain relations between concepts in a question-answering system. Weston [39] designed ten tasks and demonstrated that supervision given as feedback from simulated interlocutors enables an end-to-end memory network to better predict the next utterances. Li et al. [28] improved Weston's method with reinforcement learning and showed that asking questions leads to the better performance.

Mazumder et al. [29] proposed a method for acquiring knowledge through dialogues based on a knowledge graph. When the dialogue system cannot reply to a user question based only on the knowledge graph, it asks the user to add information related to the question and stores the knowledge extracted from the user's answer. It uses KGC to generate the answer to the initial user question, but the KGC results are not used for knowledge acquisition. Rather, our method uses the KGC results to generate system questions, so the problem we address is different from theirs.

3 KNOWLEDGE ACQUISITION THROUGH DIALOGUES USING KGC SCORES

We propose to determine which link candidate the dialogue system should ask the user to verify depending on the score of the KGC output, that is, how likely it is the predicted link is correct. By not asking about link candidates with low KGC scores, it becomes possible to avoid generating questions that may decrease the user's willingness to engage in dialogue.

An example illustrating this framework is shown in Figure 1. New link candidates are predicted from the existing knowledge graph with their scores (Step 1); the system asks the user if a predicted link is correct (Step 2) and adds the triplet corresponding to the predicted link if the user gives a positive answer. In Step 2, the

Figure 1: Knowledge acquisition dialogue systems

system asks the user about only those links whose KGC scores are high, because if the system asks about an incorrect link (e.g., “Is spaghetti marinara Chinese?”), the user might not want to continue the dialogue. The system can also ask multiple users about the result before adding such a triplet to the database [33].

We used ComplEx [38] as a KGC technique since it can represent asymmetric relations between subjects and objects and is efficient in terms of computation time. Basically any KGC techniques that can output scores for the predicted links can be used in our framework.

This study is positioned at the intersection of natural language dialogue [21, 40, 45] and interactive machine learning [1, 3, 18, 22, 26], in which the outputs of a machine learning model are visually displayed to the user [14] or are used to select informative questions for the user [13]. Our study is close to the latter and shows a promising way to obtain knowledge from the user during dialogues. In such cases, the user impression of the questions is important because the users of a dialogue system are not crowdworkers and would likely stop interacting with the system if it repeatedly asked annoying questions. As pointed out in [1], potential users have limited involvement in the training process of machine learning models. Therefore, it is crucial for dialogue-based knowledge acquisition to select reasonable questions that would not decrease their engagement.

4 KNOWLEDGE GRAPH USED IN THIS STUDY

We assume that the system has a knowledge graph in a specific domain so that it can be used in a service in that domain. The knowledge graph is initially built by hand, so it has a small amount of knowledge compared with a perfect knowledge graph in that domain. This assumption is different from those of studies on knowledge graphs that use large knowledge graphs such as Freebase [5], YAGO, [36], and WordNet [16], which are open-domain and well-crafted knowledge graphs.

We used a knowledge graph in the food and restaurant domain in our experiments. It was manually built to be used in a closed-domain chatbot in that domain [32]. The graph used in the experiments was an earlier version and different from the one in [32]. We used only the relations in the graph that were suitable to be acquired through dialogues with users. The graph contained 7,304 entities, 14 relations, and 27,914 triplets. There were 12 classes of entities.

Table 1 shows the seven main relations used in our experiments. This knowledge graph had other relations such as *alternative_name*, *area*, and *typical_side*. We did not use those because there were not enough triplets having those relations. Each relation had subject and object classes that restrict entities. Table 2 lists these classes and gives examples of them.

5 IMPROVING KGC SCORING FOR INCOMPLETE KNOWLEDGE GRAPH

Better KGC scoring leads to more appropriate questions for knowledge acquisition being selected in an early stage of development, where only an incomplete knowledge graph is available.¹ This section presents two modifications for improving KGC scoring in such a situation. Note that the modifications are not applicable to all cases. The first is not effective for almost-complete knowledge graphs with sufficient connections between entities. The second can only be applied to a knowledge graph having subject/object class information.

5.1 Two Modifications to the KGC algorithm

5.1.1 Exploiting Substrings of Entity Names. When the target knowledge graph is incomplete, the links between its entities are often insufficient. This is one of the reasons of poor KGC performance. It is desirable to augment triplets, e.g., relations between unlinked entities, with other information sources besides the links in the original knowledge graph.

¹ Our target was not an initial knowledge graph having few connections, but rather an imperfect knowledge graph having connections to which KGC can be applied.

Table 1: Relations in knowledge graph

relation	subject class	object class	example
is_a	FOOD, INGREDIENT, DRINK, CUISINE_TYPE, RESTAURANT_SHOP	FOOD, INGREDIENT, DRINK, CUISINE_TYPE, RESTAURANT_SHOP	(Camembert, is_a, cheese)
consists_of	FOOD, INGREDIENT, DRINK	FOOD, INGREDIENT, DRINK	(hamburger, consists_of, ground meat)
meal_type	FOOD, INGREDIENT, DRINK	MEAL	(pickles, meal_type, side dishes)
cuisine_type	FOOD, INGREDIENT, DRINK	CUISINE_TYPE	(spring roll, cuisine_type, Chinese)
temperature	FOOD, INGREDIENT, DRINK	TEMPERATURE	(beef stew, temperature, hot)
served_at	FOOD, INGREDIENT, DRINK	RESTAURANT_SHOP	(espresso, served_at, coffee shop)
tastes	FOOD, INGREDIENT, DRINK	TASTE	(tom yum goon, tastes, spicy)

Table 2: Entity classes

Class name	Example
FOOD	hamburger, curry and rice
INGREDIENT	salt, pasta, onion, strawberry, potato
DRINK	coffee, beer, whiskey, water
CUISINE_TYPE	Mexican, Chinese, Sichuan, vegetarian
MEAL	main dishes, side dishes, dessert
TIME	morning, summer
RESTAURANT_SHOP	Subway, Italian restaurant
AREA_PLACE	New York, resort area
OPPORTUNITY	drinking party, Christmas party
TEMPERATURE	hot, cold, at room temperature
TASTE	spicy, sweet, salty, sour
EVA_VALUE	0,1,2,3 (degree of popularity)

We assume that entities having the same substrings are likely to be related and thus connect such entities. For this purpose, we decomposed an entity name in the original knowledge graph into substrings consisting of the strings’ first or last few characters and used them as pseudo entities. A pseudo entity can be used as an augmented triplet with their original entity having a pseudo relation. Specifically, we generated substrings with N or less characters² of each entity name and used them as pseudo entities whose names start with “subs_”. The substrings were used as new triplets with their original entities having either the “prefix_” relation or the “suffix_” relation. Then, we conducted KGC with the knowledge graph augmented by the triplets having the pseudo entities and relations.

Figure 2 shows an example when $N = 2$. The two food names “mo-mi-ji-don”³ and “oya-ko-don” are in the knowledge graph, and there are originally no connections between them. Pseudo entities, such as subs_mo, subs_momi, are generated, and pseudo relations of suffix_1, suffix_2, prefix_1, and prefix_2 are also generated. As a result, the two original entities are connected via the pseudo entity “subs_don” with the “suffix_1” relation.

The use of substrings can improve KGC performance because a connection between the two original entities is created and KGC can induce the two entities to have similar properties through the created connection. This type of induction is impossible as long as the two entities have no connection between them, which is

Hyphens indicate delimiters. Unit is Japanese character.

Figure 2: Example of “mo-mi-ji-don” and “oya-ko-don” when $N = 2$

often the case when the knowledge graph is incomplete. This augmentation can improve performance as long as the entities having the same substring have similar properties and the connections between such entities have not been exhaustively given yet.

By enhancing this graph with an expansion with $N = 5$, the resulting graph contained 32,624 entities, 24 relations, and 90,318 triplets. The expansion using substrings of entity names is denoted as “Sub”.

5.1.2 Limiting Negative Sampling. Knowledge graphs generally hold only true facts represented as a triplet $(s, r, o) = 1$, and KGC models are trained by regarding the other possible triplets as false facts on the basis of the local closed world assumption [15]. Specifically, negative examples are generated in the training by randomly replacing one element of triplets representing the true facts. However, especially when the knowledge graph is incomplete and the amount of training data is limited, appropriate training samples are required to make the training process efficient. When the negative examples are randomly generated, a triplet whose relation and object do not match at all is often generated, e.g., (Sushi, temperature, sweet). Such negative examples do not contribute to improving KGC performance, because they are far from positive examples; on the other hand, negative examples can help the KGC model learn to discriminate between positives and negatives that are very similar to the positives [24].

We thus put a constraint on object entities that are replaced when generating negative examples. Specifically, the replaced object entity o has to be in the object classes of relation r , as listed in Table 1. This constraint prevents a negative example that never co-occurs with its relation, which makes the training efficient when the amount of training data is limited. We denote this limitation (negative sample filtering) as “NF”.

² We used Japanese characters as the unit.

³ Hyphens indicate delimiters between Japanese characters.

Table 3: Prediction accuracies in validation set per number of negative samples (%)

	10	30	50
Baseline	20.4	23.4	22.8
NF	20.7	23.1	23.0
Sub	29.9	33.4	33.7
Sub+NF	34.0	35.3	34.0

5.2 Experimental Results

To see if the two modifications improve KGC performance, we conducted an objective evaluation. We compared vanilla ComplEx and ComplEx with the two modifications: “Sub” and “NF”. “Sub+NF” denotes the case in which the both modifications were applied.

First, we looked at *Hits@k*, which is a commonly used metric and indicates the link prediction accuracy. To calculate it, the set of triplets in the graph was first divided into training, validation, and test sets. The embedding was learned from the training set, and a hyperparameter was optimized with the validation set. Then, for each triplet $(\hat{s}, \hat{r}, \hat{o}) = 1$ in the test set, the score for $(\hat{s}, \hat{r}, \hat{o})$ was calculated from the embedding for any entity o belonging to the classes that $\hat{r}$ takes as objects. Then, for all o , $(\hat{s}, \hat{r}, o)$ were sorted by their score. *Hits@k* is the proportion of $(\hat{s}, \hat{r}, \hat{o})$ appearing in the top- k triplets. For example, for a triplet (ice cream, tastes, sweet) in a test set, for each o in the TASTE class, the score for (ice cream, tastes, o) is calculated and sorted to check if (ice cream, tastes, sweet) is among the top- k triplets.

We computed Hits@1 for the seven main relations, namely, *is_a*, *consists_of*, *meal_type*, *cuisine_type*, *temperature*, *served_at*, and *tastes* by 10-fold cross validation. We selected these relations because there were more triplets in the target knowledge graph. We divided up the triplets into ten folds so that each fold contained almost the same number of triplets with each relation.

The training data was embedded using ComplEx [38]. We used logistic regression and the stochastic gradient descent algorithm with AdaGrad for training. The learning rate was 0.5 and the regularization factor was 0.01. We optimized the number of negative samples as one of the hyperparameters from {10, 30, 50} with the validation set in 10-fold cross validation. Table 3 lists the prediction accuracies (Hits@1) in the validation set. As a result, the best accuracies were obtained when the numbers of negative samples were 30 for Baseline, NF, and Sub+NF and 50 for Sub. We used these values in the following experiments. To limit the entities in the negative sampling, we used the classes for each relation as subjects and objects, as listed in Table 1.

Table 4 lists the Hits@1 values for each of the seven relations in the test set. The accuracies of several relations such as *temperature* and *tastes* were higher than other relations. A reason for the high accuracy was that they had restricted object classes, as in Table 1.

Table 4 also shows that the model trained with Sub+NF outperformed those of Baseline, NF, and Sub under the All condition. Specifically, the results indicate that the Sub modification improved accuracy, which was determined by comparing the differences between “Baseline” and “Sub” and between “NF” and “Sub+NF”. That is, expanding the graph structure by using substrings improved the

Table 4: Link prediction performance (Hits@1) (%)

Relation	# of triplets	Baseline	NF	Sub	Sub+NF
<i>is_a</i>	5,317	13.3	13.0	34.1	35.6
<i>consists_of</i>	3,644	7.4	6.3	16.4	14.5
<i>meal_type</i>	2,096	18.1	18.1	18.1	18.1
<i>cuisine_type</i>	1,832	50.0	49.9	55.3	59.9
<i>temperature</i>	1,141	63.0	67.4	65.5	69.1
<i>served_at</i>	821	51.2	51.4	56.8	57.7
<i>tastes</i>	560	58.2	61.4	58.8	72.0
All		24.2	24.4	34.7	36.1

accuracy for such an incomplete knowledge graph. On the other hand, the improvement due to NF was limited. This was because the class restriction was not so strong, especially for *is_a* and *consists_of* that take several classes with many entities as their subject and object.

5.3 KGC Score Distribution

We investigated how well the KGC scores for the predicted links represent the probability of their correctness. If triplets with higher KGC scores tend to be more correct, we can select correct ones from the predicted candidates. This would mean the KGC score is reliable.

We set various thresholds for the KGC scores to investigate the score’s reliability. We then counted the numbers of output triplets whose scores were larger than the thresholds and calculated the precision rates for those, i.e., the ratio of the number of correctly predicted triplets to that of output triplets. The precision rates and numbers of output triplets obtained with the Baseline and Sub+NF models are plotted for various thresholds of the KGC score. The precision rates were used here to show the correctness of triplets with higher KGC scores, with which the system will make questions. Note that we started calculating the precision rates when the number of output triplets exceeded 20 in the parts with the higher thresholds, because the rates were unstable when the number of output triplets (i.e., the denominator of the precision rate) was small.

The results for the seven relations are illustrated in Figure 3. When the thresholds for the KGC score were set higher, the numbers of output triplets decreased and the precision rates became higher in general. Note that, in the right part of the graphs, the precision rates were calculated from a small number of the output triplets. Therefore, the precision rates in that part are affected by small fluctuations in the outputs.⁴ The link prediction for *is_a* was imprecise⁵ as shown in Table 1, and thus, the KGC score’s distribution was different from the other relations. This result suggests it is not good to use triplets with higher KGC scores for *is_a*. Apart from *is_a*, the results indicate that ratios of correct results are larger

⁴This is why we displayed PR curves in Figure 4. This figure shows the overall trend of the KGC score distributions and the reliability of the scores with various thresholds.

⁵The link prediction for *is_a* was imprecise because the relation is not limited to a simple one-to-one mapping. For example, the triplet (mapo doufu, *is_a*, Sichuan) is correct, while another triplet with the same subject and relation (mapo doufu, *is_a*, Chinese) is also correct. This problem also affects *consists_of*, where, for example, both (apple pie, *consists_of*, apple) and (apple pie, *consists_of*, wheat) are correct.

Figure 3: Precision rates and output counts over thresholds

in the parts with higher KGC scores. This indicates that the score is helpful for selecting correct triplets.

Figure 3 also shows that the precision rates of the Sub+NF model were higher than those of the Baseline model for various thresholds, except in the case of the *meal_type*. These results also indicated that the scores with Sub+NF were more helpful for selecting correct triplets than those with Baseline.

We also investigated the reliability of KGC scores by plotting precision-recall (PR) curves. Figure 4 illustrates the curves for relations of *is_a*, *cuisine_type*, and *taste*, which were plotted by also

changing thresholds for the KGC scores of Baseline and Sub+NF. Although the curve stayed low for *is_a* because the link prediction for this relation was not accurate, the curves with the Sub+NF model were consistently above those of the Baseline in all cases, except *is_a*. This result also indicated that the KGC scores with the Sub+NF model performed better.

The above comparison indicates the Sub+NF model was better than the Baseline for almost all threshold values. This result suggests that better distributions of the KGC scores, shown by the PR

Figure 4: Precision and recall curves for *is_a*, *cuisine_type*, and *taste*

curves, were obtained by improvement in the KGC performance in terms of Hits@1.

6 USER STUDY OF OBTAINED QUESTIONS

We examined what kinds of questions are obtained when we used the whole knowledge graph as the training data of KGC. The obtained questions were about “truly unknown” knowledge in the original knowledge graph, while the experiments discussed in Section 5 were conducted with cross validation. In the cross validation setting, certain sets of triplets in the target knowledge graph were assumed to be unknown and their object entities were predicted to quantitatively investigate the KGC performance with data having reference labels.

We verified the questions via crowdsourcing. The results revealed how correctly the truly unknown triplets were predicted and showed that Sub+NF can select more correct questions, which would enhance user engagement in a more realistic situation.

6.1 Procedure

The experimental procedure consisted of four steps.

- (1) Conduct KGC with the Baseline and Sub+NF models, both of which were trained with the whole knowledge graph.
- (2) Select the top-20 predicted triplets sorted by the KGC scores derived by each of the two models.
- (3) Generate questions by applying hand-crafted templates to the triplets.
- (4) Ask 200 crowdworkers to provide their impressions on a 5-point scale about each question.

In Step (1), we searched all possible combinations for each relation r by replacing subjects and objects with any entity, that is, $(*, r, *)$. Note that the experiment discussed in Section 5.2 involved searching for all objects when a subject and relation pair in the test set was given, which is a standard methodology in KGC studies. We did so because we wanted to find all possible unknown triplets that are correct but are not contained in the original knowledge graph. After the search, triplets $(\hat{s}, \hat{r}, \hat{o})$ such that (i) its object $\hat{o}$ never appears as an object of the relation $\hat{r}$ in the original knowledge graph and (ii) if $(\hat{o}, \hat{r}, \hat{s})$ exists (i.e., a triplet where $\hat{s}$ and $\hat{o}$ are replaced) in the original knowledge graph, were removed from the output

Table 5: Choices of crowdworker

	Labels	Examples
+2	Obviously correct	Is curry a kind of food?
+1	Correct	Is ice cream sweet?
0	Partially correct	Is the soba cold?
-1	Incorrect	Is ice cream bitter?
-2	Obviously incorrect	Is flour used to make fish?
Idk	I don't know	

because they are rarely correct. We also removed a few entities that were invalid as a single-sentence question, e.g., ingredient-specific dishes as a cuisine.

The KGC models were trained in the same way as the experiments in Section 5. More specifically, the embedded representation was trained using ComplEx [38]. We used logistic regression and the stochastic gradient descent algorithm with AdaGrad for training. The learning rate was 0.5, the regularization factor was 0.01, the number of negative samples was 30, and the number of dimensions of the embedded vector was 200. The parameter of the substring of the entity names was 5.

In Step (2), the unknown triplets were predicted for six relations: *is_a*, *consists_of*, *cuisine_type*, *temperature*, *served_at*, and *tastes*. We did not include *meal_type* in the following crowdsourcing because the performances of *meal_type* for the two models were the same in the validation-based evaluation shown in Table 4. We calculated the KGC scores of all combinations for subjects and objects for the relations and sorted the predicted triplets based on these scores for each relation. We then extracted the top-20 triplets for each relation and obtained $20 \times 6 = 120$ triplets for each model.

In Step (3), question sentences were generated by applying hand-crafted templates to the predicted triplets. Templates in Japanese were manually prepared for each relation. For example, the template of the *is_a* relation is “*subject* is a kind of *object*?” Their expressions were basically fixed per relation, except that “eat” was replaced with “drink” when the class of the target entity was DRINK.

In Step (4), 240 questions (120 each for Baseline and Sub+NF) were displayed to 200 crowdworkers, who were recruited from

Table 6: Frequency of annotated labels per relation for questions obtained with Baseline

	+2	+1	0	-1	-2	Idk	Total
is_a	202	322	248	510	648	70	2000
consists_of	273	548	352	405	354	68	2000
cuisine_type	72	230	294	747	566	91	2000
temperature	157	507	1062	219	50	5	2000
served_at	195	403	683	400	270	49	2000
tastes	41	441	1118	254	125	21	2000
Total	940	2451	3757	2535	2013	304	12000

the general public via a crowdsourcing platform.⁶ We divided the questions into two equivalent groups of 120 questions (60 each for the two models) to prevent the workers from becoming overloaded with questions. We then displayed the questions in one group in random order to each worker; 100 workers rated the correctness of the questions in one group, and the remaining 100 workers rated those in the other group. That is, all 240 questions were rated by 100 workers. The workers were asked to rate each displayed question by selecting from the choices in Table 5. There were five numerical choices, ranging from “+2: Obviously correct” to “-2: Obviously incorrect”, and a non-numerical choice indicating “I don’t know” (labeled “Idk”). We asked the workers to look up the answers using a web search when they did not know it or they did not understand the question. If a worker selected “I don’t know” for more than 25% of the questions, we discarded that worker’s results and asked another worker to rate.

6.2 Results

We obtained in total 24,000 annotated labels, consisting the five numerical ratings and “Idk”, to the 240 questions. The annotation results are shown in Tables 6 and 7, which indicate the frequency of the annotated labels for questions obtained with Baseline and Sub+NF. As in Table 4, *temperature*, *served_at*, and *tastes*, which had restricted object classes, obtained relatively more positive ratings. We also calculated the agreement of the annotation results in terms of Krippendorff’s alpha [25] by assuming the labels other than the Idk labels were numeric values.⁷ The Krippendorff’s alpha with the interval metric differences was 0.557 for one group of 120 questions with 11,730 samples and 0.549 for the other group with 11,785 samples. This shows that there was moderate agreement [4] among the annotation results.⁸

Table 8 shows the ratios of each label for the two models. The ratios in the table summarizes the results in Tables 6 and 7. Under the assumption that the +2 and +1 annotation results are correct, the accuracy of Sub+NF is 44.0%, which is much higher than 36.1% in Table 4. The higher accuracy in Table 8 is because we selected the top-20 questions sorted on the basis of the KGC scores. The results indicate that the KGC score is helpful in selecting more correct

Table 7: Frequency of annotated labels per relation for questions obtained with Sub+NF

	+2	+1	0	-1	-2	Idk	Total
is_a	381	498	227	507	334	53	2000
consists_of	453	549	414	295	252	37	2000
cuisine_type	186	385	419	456	507	47	2000
temperature	224	402	1199	116	51	8	2000
served_at	312	676	624	236	129	23	2000
tastes	312	898	650	98	29	13	2000
Total	1868	3408	3533	1708	1302	181	12000

Table 8: Ratios of each label per model (%)

	+2	+1	0	-1	-2	Idk
Baseline	7.8	20.4	31.3	21.1	16.8	2.5
Sub+NF	15.6	28.4	29.4	14.2	10.9	1.5

questions, which will not diminish the user’s willingness to engage in dialogue with the system.

Table 8 shows that the accuracies of Baseline and Sub+NF were 28.2% and 44.0% under the assumption that the +2 and +1 annotation results were correct. The difference in accuracies was statistically significant from a z-test ($Z = 20.6$; $p < 0.01$). The results also indicate that Sub+NF selected more correct questions about the truly unknown triplets. Several examples are given in Section 6.3.

Table 8 also shows that the ratios of “+2 Obviously correct” were 7.8% for Baseline and 15.6% for Sub+NF, while the ratios of “+1 Correct” increased from 20.4% for Baseline to 28.4% for Sub+NF. The results indicate that Sub+NF also selected more obviously correct questions, which were derived by using substrings of entity names. For example, a triplet (tomato juice, is_a, juice), which was missing from the original knowledge graph, was predicted correctly based on the KGC scores and was selected as a question: “Is tomato juice a kind of juice?” The triplets received a high score because the two entity names share the substring “juice” (in Japanese characters). Although it will be useful to add the correct triplet to the knowledge graph for future use, such questions may be perceived as annoying although their content is correct. The system should avoid asking such a question several times. In practical cases when the system would actually add a triplet to the system knowledge, the system would ask a question about the triplet to several human users [33] and obtain answers indicating it is correct. In such a case, a strategy can be taken to reduce the number of questions having such a high KGC score, which may be obviously correct. New strategies to avoid obviously correct questions will be left for future consideration.

The crowdsourcing results indicate that Sub+NF improved KGC performance for truly unknown triplets in the target knowledge graph. It also results in better questions that do not diminish the user’s willingness to engage in dialogue with the system.

6.3 Examples

Example questions are listed in Table 9. The column “Label Ave.” denotes the average annotation values of the labels other than the “Idk” labels. We selected several questions as examples indicating

⁶ We used the platform of CrowdWorks, Inc. <https://crowdworks.co.jp/>

⁷ Because the labels were displayed to the crowdworkers together with numeric values, we supposed that the labels could be regarded as composing an interval scale and calculated the agreement among the labels.

⁸ The labels cannot be determined uniquely because they were subjective. For example, some crowdworkers indicated that the question “Are vegetables hot?” is partially correct (0), while other crowdworkers indicated that it is incorrect (−1).

Table 9: Example questions

Question	Label Ave.	Model
Is curry rice served at a curry shop?	1.86	Baseline
Is a frozen drink cold?	1.73	Sub+NF
Does mango mousse tart taste sweet?	1.10	Sub+NF
Do sesame dumplings taste sweet?	0.93	Sub+NF
Is white sauce used to make chicken rice doria?	0.93	Baseline
Is Calpis Soda cold?	0.89	Sub+NF
Is milk cold?	0.09	Sub+NF
Does red wine taste refreshing?	0.01	Baseline
Are vegetables hot?	-0.88	Sub+NF
Is gazpacho drunk at a German restaurant?	-1.03	Baseline
Is a Latin American restaurant a kind of Italian restaurant?	-1.71	Baseline
Is Thai curry French food?	-1.82	Sub+NF

each of the five labels from -2 to +2. The distributions of the annotated labels are in Table 8. Note that all the triplets used to make the questions were unknown in the original knowledge graph.

There were reasonable questions whose label averages were around +1, e.g., “Does mango mousse tart taste sweet?” and “Is white sauce used to make chicken rice doria?”. There were several questions that many crowdworkers judged as “obviously correct (+2)”, e.g., “Is a frozen drink cold?”. The questions with labels close to 0 were not always correct. Milk can be hot or cold, and red wine has a variety of tastes. Incorrect questions with negative annotation labels were also selected because of the incompleteness of the target knowledge graph and lack of triplets for training.

7 CONCLUDING REMARKS

Knowledge bases are crucial for various information systems. It would be very helpful if a natural language dialogue interfaces could enhance their knowledge bases by asking appropriate questions to human users. However, developing a means of obtaining appropriate questions from a given knowledge base is a challenge, and the questions should not diminish the human user’s willingness to engage in dialogue with the system. In this study, we used a KGC technique and its scores to select such questions.

In particular, our study dealt with the following. First, when KGC performance is low, too many incorrect questions are selected, which may diminish the user’s willingness to engage in dialogue with the system. To deal with this problem, We applied two modifications during training to the incomplete knowledge graph; the Sub modification uses substrings of entity names and NF modification imposes limitations on the negative sampling. As a result, KGC performance improved, especially with Sub. Note that the Sub modification worked well in the food and restaurant domain, but its applicability is limited to domains where the entities having the same substring have similar properties and the connections between such entities have not been given exhaustively.

Second, we showed that better questions were selected using the improved KGC scores with the Sub+NF model. It is crucial to consider the users’ impressions of the questions because the user of the dialogue interface is not an “oracle” in an active learning setting [12] but rather someone who is using the interface voluntarily.

Therefore, from that viewpoint, we attempted to select more correct questions.

Our future work will include a further user study in which the proposed framework will be implemented into a dialogue system. An interesting issue is the way to enable the system to make the user enjoy a chitchat conversation with it and at the same time acquire knowledge during the dialogue. The key aspect is that system’s questions should not be annoying. The applicability of the dialogue approach depends on whether the user can answer the system’s questions. As well, setting an appropriate target domain in which the user can answer the questions and enjoy dialogues with the system is an important future issue.⁹ Another issue is the system’s strategy to change the forms of its questions according to the KGC scores and to reduce the number of verification to multiple users [33] before acquiring a triplet as knowledge having a high KGC score, which may be obviously correct. Another issue is to determine how well the triplets obtained with the selected questions improve KGC performance. Questions that the user indicates as incorrect would be able to be used as negative samples. This problem can be formulated as a kind of interactive machine learning [1], which is an interesting future direction of this study. Yet another issue is to verify whether the improvement derived by the use of substrings in KGC depends on the language. Experiments with another incomplete knowledge base will be among our future work.

ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Number JP19H04171.

REFERENCES

- [1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (December 2014). <https://doi.org/10.1609/aimag.v35i4.2513>
- [2] Renzo Angles and Claudio Gutierrez. 2008. Survey of Graph Database Models. *Comput. Surveys* 40, 1 (2008), 1–39.
- [3] Dustin Arendt, Emily Saldanha, Ryan Wesslen, Svitlana Volkova, and Wenwen Dou. 2019. Towards Rapid Interactive Machine Learning: Evaluating Tradeoffs of Classification without Representation. In *Proc. International Conference on Intelligent User Interfaces (IUI)* (Marina del Ray, California) (IUI ’19). Association

⁹ We selected the food and restaurant domain for the experiment because we thought that crowdworkers would easily understand the system’s questions in this domain.

- for Computing Machinery, New York, NY, USA, 591–602. <https://doi.org/10.1145/3301275.3302280>
- [4] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (Dec. 2008), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- [5] 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 (Vancouver, Canada)*. Association for Computing Machinery, 1247–1250. <https://doi.org/10.1145/1376616.1376746>
- [6] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning Structured Embeddings of Knowledge Bases. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (San Francisco, California). AAAI Press, 301–306.
- [7] Andrew Carlson, Justin Betteridge, Bryan Kiesel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an Architecture for Never-Ending Language Learning. In *Proc. Conference on Artificial Intelligence (AAAI)*. <http://rtw.ml.cmu.edu/papers/carlson-aaai10.pdf>
- [8] Asli Celikyilmaz, Dilek Hakkani-Tür, Panupong Pasupat, and Ruhi Sarikaya. 2015. Enriching Word Embeddings Using Knowledge Graph for Semantic Tagging in Conversational Dialog Systems. In *Proc. Conference on Artificial Intelligence (AAAI)*. <https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/viewFile/10333/10034>
- [9] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <https://www.aclweb.org/anthology/D19-1189/>
- [10] Yun-Nung Chen, William Yang Wang, Anatole Gershman, and Alexander Rudnicky. 2015. Matrix Factorization with Knowledge Graph Propagation for Unsupervised Spoken Language Understanding. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*. <https://www.aclweb.org/anthology/P15-1047>
- [11] Zhiyuan Chen and Bing Liu. 2018. *Lifelong Machine Learning, Second Edition*. Morgan & Claypool Publishers. <https://doi.org/10.2200/S00832ED1V01Y201802AIM037>
- [12] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active Learning with Statistical Models. *Journal of Artificial Intelligence Research* 4, 1 (March 1996), 129–145.
- [13] Pedram Daei, Tomi Peltola, Marta Soare, and Samuel Kaski. 2017. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Machine Learning* 106, 9-10 (2017), 1599–1620. <https://doi.org/10.1007/s10994-017-5651-7>
- [14] Pedram Daei, Tomi Peltola, Aki Vehtari, and Samuel Kaski. 2018. User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction. In *Proc. International Conference on Intelligent User Interfaces (IUI)* (Tokyo, Japan) (IUI '18). Association for Computing Machinery, New York, NY, USA, 305–310. <https://doi.org/10.1145/3172944.3172989>
- [15] Xin Dong, Evgeniy Gabrilovich, Jeremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 601–610. <https://doi.org/10.1145/2623330.2623623>
- [16] Christiane Fellbaum (Ed.). 1998. *WordNet: an electronic lexical database*. MIT Press.
- [17] Edward Filisko and Stephanie Seneff. 2005. Developing city name acquisition strategies in spoken dialogue systems via user simulation. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (SIGDIAL 2005)*. 144–155.
- [18] Marco Gillies, Andrea Kleinsmith, and Harry Brenton. 2015. Applying the CASSM Framework to Improving End User Debugging of Interactive Machine Learning. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, Oliver Brdiczka, Polo Chau, Giuseppe Carenini, Shimei Pan, and Per Ola Kristensson (Eds.). ACM, 181–185. <https://doi.org/10.1145/2678025.2701373>
- [19] Larry Heck and Hongzhao Huang. 2014. Deep learning of knowledge graph embeddings for semantic parsing of twitter dialogs. In *Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. <https://ieeexplore.ieee.org/abstract/document/7032187>
- [20] Ben Hixon, Peter Clark, and Hannaneh Hajishirzi. 2015. Learning Knowledge Graphs for Question Answering through Conversational Dialog. In *Proc. North American Chapter of Association for Computational Linguistics (NAACL)* (Denver, Colorado). 851–861. <https://doi.org/10.3115/v1/N15-1086>
- [21] Forrest Huang, Eldon Schoop, David Ha, and John F. Canny. 2020. Scones: towards conversational authoring of sketches. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, Fabio Paternò, Nuria Oliver, Cristina Conati, Lucio Davide Spano, and Nava Tintarev (Eds.). ACM, 313–323. <https://doi.org/10.1145/3377325.3377485>
- [22] Biye Jiang and John F. Canny. 2017. Interactive Machine Learning via a GPU-accelerated Toolkit. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu (Eds.). ACM, 535–546. <https://doi.org/10.1145/3025171.3025172>
- [23] Kazunori Komatani and Mikio Nakano. 2020. User Impressions of Questions to Acquire Lexical Knowledge. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Association for Computing Machinery, 1st virtual meeting, 147–156. <https://www.aclweb.org/anthology/2020.sigdial-1.19>
- [24] Bhushan Kotnis and Vivi Nastase. 2017. Analysis of the Impact of Negative Sampling on Link Prediction in Knowledge Graphs. *arXiv:1708.06816* (2017). [arXiv:1708.06816](https://arxiv.org/abs/1708.06816)
- [25] Klaus Krippendorff. 2011. Computing Krippendorff’s alpha-reliability. https://repository.upenn.edu/asc_papers/43/.
- [26] Todd Kulesza, Margaret M. Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, Oliver Brdiczka, Polo Chau, Giuseppe Carenini, Shimei Pan, and Per Ola Kristensson (Eds.). ACM, 126–137. <https://doi.org/10.1145/2678025.2701399>
- [27] Ni Lao, Tom Mitchell, and William W. Cohen. 2011. Random Walk Inference and Learning in a Large Scale Knowledge Base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 529–539. <https://www.aclweb.org/anthology/D11-1049>
- [28] Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc’Aurelio Ranzato, and Jason Weston. 2017. Learning through Dialogue Interactions by Asking Questions. In *Proc. International Conference on Learning Representations (ICLR)*. <https://openreview.net/pdf?id=rkE8pVcle>
- [29] Sahisnu Mazumder, Bing Liu, Shuai Wang, and Nianzu Ma. 2019. Lifelong and Interactive Learning of Factual Knowledge in Dialogues. In *Proceedings of Annual Meeting of the Special Interest Group on Discourse and Dialogue*.
- [30] Helen Meng, P. C. Ching, Shuk Fong Chan, Yee Fong Wong, and Cheong Chat Chan. 2004. ISIS: An Adaptive, Trilingual Conversational System With Interleaving Interaction and Delegation Dialogs. *ACM Transactions on Computer-Human Interaction* 11, 3 (2004), 268–299.
- [31] Tom M. Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kiesel, Jayant Krishnamurthy, Ni Lao nad Kathryn Mazaitis, Tahir Mohamed, Ndapa Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, and Malcolm Greavesand Joel Welling. 2015. Never-Ending Learning. In *Proc. Conference on Artificial Intelligence (AAAI)*. <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/10049>
- [32] Mikio Nakano and Kazunori Komatani. 2020. A framework for building closed-domain chat dialogue systems. *Knowledge-Based Systems* 204 (2020), 106212. <https://doi.org/10.1016/j.knsys.2020.106212>
- [33] Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano, and Kazunori Komatani. 2017. Lexical Acquisition through Implicit Confirmations over Multiple Dialogues. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 50–59. <https://doi.org/10.18653/v1/W17-5507>
- [34] Tsugumi Otsuka, Kazunori Komatani, Satoshi Sato, and Mikio Nakano. 2013. Generating More Specific Questions for Acquiring Attributes of Unknown Concepts from Users. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 70–77.
- [35] Aasish Pappu and Alexander Rudnicky. 2014. Knowledge Acquisition Strategies for Goal-Oriented Dialog Systems. In *Proc. Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 194–198. <https://doi.org/10.3115/v1/W14-4326>
- [36] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web (Banff, Alberta, Canada)*. Association for Computing Machinery, 697–706. <https://doi.org/10.1145/1242572.1242667>
- [37] Yasuhiro Takahashi, Kohji Dohsaka, and Kiyooki Aikawa. 2002. An efficient dialogue control method using decision tree-based estimation of out-of-vocabulary word attributes. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-96)*. 813–816.
- [38] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. 2071–2080.
- [39] Jason Weston. 2016. Dialog-based Language Learning. In *Proc. International Conference on Neural Information Processing Systems (NIPS)* (Barcelona, Spain). 829–837. <http://dl.acm.org/citation.cfm?id=3157096.3157189>
- [40] Ziang Xiao, Michelle X. Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: using a conversational agent to understand individuals and help teaming. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, Wai-Tat Fu, Shimei Pan, Oliver Brdiczka, Polo Chau, and Gaelle Calvary (Eds.). ACM, 437–447. <https://doi.org/10.1145/3301275.3302264>
- [41] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proc. Conference on Artificial*

- Intelligence (AAAI)*. <https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563>
- [42] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting End-to-End Dialogue Systems with Commonsense Knowledge. In *Proc. Conference on Artificial Intelligence (AAAI)*. <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16573/0>
 - [43] Jie Zeng and Yukiko I. Nakano. 2020. Exploiting a Large-Scale Knowledge Graph for Question Generation in Food Preference Interview Systems. In *Proc. International Conference on Intelligent User Interfaces (IUI)* (Cagliari, Italy). 53–54. <https://doi.org/10.1145/3379336.3381504>
 - [44] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*. <https://www.aclweb.org/anthology/2020.acl-main.184/>
 - [45] Yunfeng Zhang, Q. Vera Liao, and Biplav Srivastava. 2018. Towards an Optimal Dialog Strategy for Information Retrieval Using Both Open- and Close-Ended Questions. In *Proc. International Conference on Intelligent User Interfaces (IUI)* (Tokyo, Japan) (*IUI '18*). Association for Computing Machinery, New York, NY, USA, 365?369. <https://doi.org/10.1145/3172944.3172998>
 - [46] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*. <https://www.ijcai.org/Proceedings/2018/0643.pdf>
 - [47] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1006–1024. <https://dl.acm.org/doi/abs/10.1145/3394486.3403143>