Recent Advances in Neural Question Generation

Liangming Pan, Wenqiang Lei, Tat-Seng Chua and Min-Yen Kan
School of Computing
National University of Singapore
Singapore 117417
{e0272310,wenql,kanmy,chuats}comp.nus.edu.sg

Abstract

Emerging research in Neural Question Generation (NQG) has started to integrate a larger variety of inputs, and generating questions requiring higher levels of cognition. These trends point to NQG as a bellwether for NLP, about how human intelligence embodies the skills of curiosity and integration.

We present a comprehensive survey of neural question generation, examining the corpora, methodologies, and evaluation methods. From this, we elaborate on what we see as emerging on NQG's trend: in terms of the learning paradigms, input modalities, and cognitive levels considered by NQG. We end by pointing out the potential directions ahead.

1 Introduction

Question Generation (QG) concerns the task of "automatically generating questions from various inputs such as raw text, database, or semantic representation" (Rus et al., 2008). People have the ability to ask rich, creative, and revealing questions (Rothe et al., 2017); *e.g.*, asking *Why did Gollum betray his master Frodo Baggins?* after reading the fantasy novel *The Lord of the Rings*. How can machines be endowed with the ability to ask relevant and to-the-point questions, given various inputs? This is a challenging, complementary task to Question Answering (QA). Both QA and QG require an in-depth understanding of the input source and the ability to reason over relevant contexts. But beyond understanding, QG additionally integrates the challenges of Natural Language Generation (NLG), *i.e.*, generating grammatically and semantically correct questions.

QG is of practical importance: in education, forming good questions are crucial for evaluating students knowledge and stimulating self-learning. QG can generate assessments for course materials (Heilman and Smith, 2010) or be used as a component in adaptive, intelligent tutoring systems (Lindberg et al., 2013). In dialog systems, fluent QG is an important skill for chatbots, *e.g.*, in initiating conversations or obtaining specific information from human users. QA and reading comprehension also benefit from QG, by reducing the needed human labor for creating large-scale datasets. We can say that traditional QG mainly focused on generating factoid questions from a single sentence or a paragraph, spurred by a series of workshops during 2008–2012 (Rus and Lester, 2009; Rus et al., 2010, 2011, 2012).

Recently, driven by advances in deep learning, QG research has also begun to utilize "neural" techniques, to develop end-to-end neural models to generate deeper questions (Chen et al., 2018) and to pursue broader applications (Serban et al., 2016; Mostafazadeh et al., 2016).

While there have been considerable advances made in NQG, the area lacks a comprehensive survey. This paper fills this gap by presenting a systematic survey on recent development of NQG, focusing on three emergent trends that deep learning has brought in QG: (1) the change of learning paradigm, (2) the broadening of the input spectrum, and (3) the generation of deep questions.

2 Fundamental Aspects of NQG

For the sake of clean exposition, we first provide a broad overview of QG by conceptualizing the problem from the perspective of the three introduced aspects: (1) its learning paradigm, (2) its input modalities, and (3) the cognitive level it involves. This combines past research with recent trends, providing insights on how NQG connects to traditional QG research.

2.1 Learning Paradigm

QG research traditionally considers two fundamental aspects in question asking: "What to ask" and "How to ask". A typical QG task considers the identification of the important aspects to ask about ("what to ask"), and learning to realize such identified aspects as natural language ("how to ask"). Deciding what to ask is a form of machine understanding: a machine needs to capture important information dependent on the target application, akin to automatic summarization. Learning how to ask, however, focuses on aspects of the language quality such as grammatical correctness, semantically preciseness and language flexibility.

Past research took a reductionist approach, separately considering these two problems of "what" and "how" via *content selection* and *question construction*. Given a sentence or a paragraph as input, content selection selects a particular salient topic worthwhile to ask about and determines the question type (*What*, *When*, *Who*, etc.). Approaches either take a syntactic (Gates, 2008; Liu et al., 2010; Heilman, 2011) or semantic (Yao et al., 2012; Lindberg et al., 2013; Mazidi and Nielsen, 2014; Chali and Hasan, 2015) tack, both starting by applying syntactic or semantic parsing, respectively, to obtain intermediate symbolic representations. Question construction then converts intermediate representations to a natural language question, taking either a *tranformation-* or *template-based* approach. The former (Ali et al., 2010; Pal et al., 2010; Heilman, 2011) rearranges the surface form of the input sentence to produce the question; the latter (Chen and Mostow, 2009; Liu et al., 2012; Rokhlenko and Szpektor, 2013) generates questions from pre-defined question templates. Unfortunately, such QG architectures are limiting, as their representation is confined to the variety of intermediate representations, transformation rules or templates.

In contrast, neural models motivate an end-to-end architectures. Deep learned frameworks contrast with the reductionist approach, admitting approaches that jointly optimize for both the "what" and "how" in an unified framework. The majority of current NQG models follow the sequence-to-sequence (Seq2Seq) framework that use a unified representation and joint learning of content selection (via the encoder) and question construction (via the decoder). In this framework, traditional parsing-based content selection has been replaced by more flexible approaches such as attention (Bahdanau et al., 2014) and copying mechanism (Gülçehre et al., 2016). Question construction has become completely data-driven, requiring far less labor compared to transformation rules, enabling better language flexibility compared to question templates.

However, unlike other Seq2Seq learning NLG tasks, such as Machine Translation, Image Captioning, and Abstractive Summarization, which can be loosely regarded as learning a one-to-one mapping, generated questions can differ significantly when the intent of asking differs (*e.g.*, the target answer, the target aspect to ask about, and the question's depth). In Section 5, we summarize different NQG methodologies based on Seq2Seq framework, investigating how some of these QG-specific factors are integrated with neural models, and discussing what could be further explored. The change of learning paradigm in NQG era is also represented by multi-task learning with other NLP tasks, for which we discuss in Section 6.1.

2.2 Input Modality

Question generation is an NLG task for which the input has a wealth of possibilities depending on applications. While a host of input modalities have been considered in other NLG tasks, such as text summarization (Mani, 1999), image captioning (Vinyals et al., 2015) and table-to-text generation (Lebret et al., 2016), traditional QG mainly focused on textual inputs, especially declarative sentences, explained by the original application domains of question answering and education, which also typically featured textual inputs.

Recently, with the growth of various QA applications such as Knowledge Base Question Answering (KBQA) (Cui et al., 2017) and Visual Question Answering (VQA) (Antol et al., 2015), NQG research has also widened the spectrum of sources to include knowledge bases (Khapra et al., 2017) and images (Mostafazadeh et al., 2016). This trend is also spurred by the remarkable success of neural models in feature representation, especially on image features (Krizhevsky et al., 2012) and knowledge representations (Bordes et al., 2013). We discuss adapting NQG models to other input modalities in Section 6.2.

2.3 Cognitive Levels

Finally, we consider the required cognitive process behind question asking, a distinguishing factor for questions (Anderson et al., 2001). A typical framework that attempts to categorize the cognitive levels involved in question asking comes from Bloom's taxonomy (Bloom et al., 1984), which has undergone several revisions and currently has six cognitive levels: *Remembering*, *Understanding*, *Applying*, *Analyzing*, *Evaluating* and *Creating* (Anderson et al., 2001).

Traditional QG focuses on shallow levels of Bloom's taxonomy: typical QG research is on generating sentence-based factoid questions (*e.g.*, *Who*, *What*, *Where* questions), whose answers are simple constituents in the input sentence (Heilman and Smith, 2010; Heilman, 2011). However, a QG system achieving human cognitive level should be able to generate meaningful questions that cater to higher levels of Bloom's taxonomy (Desai et al., 2018), such as *Why*, *What-if*, and *How* questions. Traditionally, those "deep" questions are generated through shallow methods such as handcrafted templates (Liu et al., 2012; Rokhlenko and Szpektor, 2013); however, these methods lack a real understanding and reasoning over the input.

Although asking deep questions is complex, NQG's ability to generalize over voluminous data has enabled recent research to explore the comprehension and reasoning aspects of QG (Labutov et al., 2015; Rothe et al., 2017; Chen et al., 2018; Desai et al., 2018). We investigate this trend in Section 6.3, examining the limitations of current Seq2Seq model in generating deep questions, and the efforts made by existing works, indicating further directions ahead.

The rest of this paper provides a systematic survey of NQG, covering corpus and evaluation metrics before examining specific neural models.

3 Corpora

As QG can be regarded as a dual task of QA, in principle any QA dataset can be used for QG as well. However, there are at least two corpus-related factors that affect the difficulty of question generation. The first is the required **cognitive level** to answer the question, as we discussed in the previous section. Current NQG has achieved promising results on datasets consisting mainly of shallow factoid questions, such as SQuAD (Rajpurkar et al., 2016) and MS MARCO (Nguyen et al., 2016). However, the performance drops significantly on deep question datasets, such as LearningQ (Chen et al., 2018), shown in Section 6.3. The second factor is the **answer type**, i.e., the expected form of the answer, typically having four settings: (1) the answer is a text span in the passage, which is usually the case for factoid questions, (2) human-generated, abstractive answer that may not appear in the passage, usually the case for deep questions, (3) multiple choice question where question and its distractors should be jointly generated, and (4) no given answer, which requires the model to automatically learn what is worthy to ask. The design of NQG system differs accordingly.

Table 1 presents a listing of the NQG corpora grouped by their cognitive level and answer type, along with their statistics. Among them, SQuAD was used by most groups as the benchmark to evaluate their NQG models. This provides a fair comparison between different techniques. However, it raises the issue that most NQG models work on factoid questions with answer as text span, leaving other types of QG problems less investigated, such as generating deep multi-choice questions. To overcome this, a wider variety of corpora should be benchmarked against in future NQG research.

4 Evaluation Metrics

Although the datasets are commonly shared between QG and QA, it is not the case for evaluation: it is challenging to define a gold standard of proper questions to ask. Meaningful, syntactically correct, semantically sound and natural are all useful criteria, yet they are hard to quantify. Most QG systems involve *human evaluation*, commonly by randomly sampling a few hundred generated questions, and asking human annotators to rate them on a 5-point Likert scale. The average rank or the percentage of best-ranked questions are reported and used for quality marks.

As human evaluation is time-consuming, common automatic evaluation metrics for NLG, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), and ROUGE (Lin, 2004), are also widely used. However, some studies (Callison-Burch et al., 2006; Liu et al., 2016) have shown that these metrics do not correlate well with fluency, adequacy, coherence, as they essentially compute the n-gram similarity between the source sentence and the generated question. To overcome this, Nema and Khapra (2018) proposed a new metric to evaluate the "answerability" of a question by calculating the scores for several question-specific

Cognitive Level	Dataset / Contributor	Answer Type	Domain	Statistics		
				Documents	Questions	Q./Doc
Shallow	SQuAD (Rajpurkar et al., 2016)	text span	Wikipedia	20,958	97,888	4.67
	NewsQA (Trischler et al., 2017)	text span	News	12,744	119,633	9.39
Medium	MS MARCO (Nguyen et al., 2016)	human generated	Web article	1,010,916	3,563,535	3.53
	RACE (Lai et al., 2017)	multiple choice	Education	27,933	72,547	2.60
Deep	LearningQ (Chen et al., 2018)	no answer	Education	10,841	231,470	21.35
	NarrativeQA (Kociský et al., 2018)	human generated	Story	1,572	46,765	29.75

Table 1: NQG datasets grouped by their cognitive level and answer type, where the number of documents, the number of questions, and the average number of questions per document (Q./Doc) for each corpus are listed.

factors, including question type, content words, function words, and named entities. However, as it is newly proposed, it has not been applied to evaluate any NQG system yet.

To accurately measure what makes a good question, especially deep questions, improved evaluation schemes are required to specifically investigate the mechanism of question asking.

5 Methodology

Many current NQG models follow the Seq2Seq architecture. Under this framework, given a passage (usually a sentence) $X = (x_1, \cdots, x_n)$ and (possibly) a target answer A (a text span in the passage) as input, an NQG model aims to generate a question $Y = (y_1, \cdots, y_m)$ asking about the target answer A in the passage X, which is defined as finding the best question $\bar{Y}$ that maximizes the conditional likelihood given the passage X and the answer A:

$$\bar{Y} = \arg\max_{Y} P(Y|X, A) \tag{1}$$

$$= \arg\max_{Y} \sum_{t=1}^{m} P(y_t|X, A, y_{<t}) \tag{2}$$

Du et al. (2017) pioneered the first NQG model using an attention Seq2Seq model (Bahdanau et al., 2014), which feeds a sentence into an RNN-based encoder, and generate a question about the sentence through a decoder. The attention mechanism is applied to help decoder pay attention to the most relevant parts of the input sentence while generating a question. Note that this base model does not take the target answer as input. Subsequently, neural models have adopted attention mechanism as a default (Zhou et al., 2017; Duan et al., 2017; Harrison and Walker, 2018).

Although these NQG models all share the Seq2Seq framework, they differ in the consideration of — (1) QG-specific factors (*e.g.*, answer encoding, question word generation, and paragraph-level contexts), and (2) common NLG techniques (*e.g.*, copying mechanism, linguistic features, and reinforcement learning) — discussed next.

5.1 Encoding Answers

The most commonly considered factor by current NQG systems is the target answer, which is typically taken as an additional input to guide the model in deciding which information to focus on when generating; otherwise, the NQG model tend to generate questions without specific target (*e.g.*, "What is mentioned?"). Models have solved this by either treating the answer's position as an extra input feature (Zhou et al., 2017; Zhao et al., 2018), or by encoding the answer with a separate RNN (Duan et al., 2017; Kim et al., 2019).

The first type of method augments each input word vector with an extra *answer indicator feature*, indicating whether this word is within the answer span. Zhou et al. (2017) implement this feature using the BIO tagging scheme, while Harrison and Walker (2018) directly use a binary indicator. In addition to the target answer, Sun et al. (2018) argued that the context words closer to the answer also deserve more attention from the model, since they are usually more relevant. To this end, they incorporate trainable position embeddings $(d_{p_1}, d_{p_2}, \cdots, d_{p_n})$ into the computation of attention distribution, where p_i is the relative distance between the i-th word and the answer, and d_{p_i} is the embedding of p_i. This achieved an extra BLEU-4 gain of 0.89 on SQuAD.

To generate answer-related questions, extra answer indicators explicitly emphasize the importance of answer; however, it also increases the tendency that generated questions include words from the answer, resulting in useless questions, as observed by Kim et al. (2019). For example, given the input "John Francis OHara was elected president of Notre Dame in 1934.", an improperly generated question would be "Who was elected John

Francis?”, which exposes some words in the answer. To address this, they propose to replace the answer into a special token for passage encoding, and a separate RNN is used to encode the answer. The outputs from two encoders are concatenated as inputs to the decoder. Song et al. (2018) adopted a similar idea that separately encodes passage and answer, but they instead use the multi-perspective matching between two encodings as an extra input to the decoder.

We forecast treating the passage and the target answer separately as a future trend, as it results in a more flexible model, which generalizes to the abstractive case when the answer is not a text span in the input passage. However, this inevitably increases the model complexity and difficulty in training.

5.2 Question Word Generation

Question words (*e.g.*, “when”, “how”, and “why”) also play a vital role in QG; Sun et al. (2018) observed that the mismatch between generated question words and answer type is common for current NQG systems. For example, a when-question should be triggered for answer “the end of the Mexican War” while a why-question is generated by the model. A few works (Duan et al., 2017; Sun et al., 2018) considered question word generation separately in model design.

Duan et al. (2017) proposed to first generate a question template that contains question word (*e.g.*, “how to #”, where # is the placeholder), before generating the rest of the question. To this end, they train two Seq2Seq models; the former learns to generate question templates for a given text , while the latter learns to fill the blank of template to form a complete question. Instead of a two-stage framework, Sun et al. (2018) proposed a more flexible model by introducing an additional decoding mode that generates the question word. When entering this mode, the decoder produces a question word distribution based on a restricted set of vocabulary using the answer embedding, the decoder state, and the context vector. The switch between different modes is controlled by a discrete variable produced by a learnable module of the model in each decoding step.

Determining the appropriate question word harks back to question type identification, which is correlated with the question intention, as different intents may yield different questions, even when presented with the same (passage, answer) input pair. This points to the direction of exploring question pragmatics, where external contextual information (such as intent) can inform and influence how questions should optimally be generated.

5.3 Paragraph-level Contexts

Leveraging rich paragraph-level contexts around the input text is another natural consideration to produce better questions. According to (Du et al., 2017), around 20% of questions in SQuAD require paragraph-level information to be answered. However, as input texts get longer, Seq2Seq models have a tougher time effectively utilizing relevant contexts, while avoiding irrelevant information.

To address this challenge, Zhao et al. (2018) proposed a gated self-attention encoder to refine the encoded context by fusing important information with the context’s self-representation properly, which has achieved state-of-the-art results on SQuAD. The long passage consisting of input texts and its context is first embedded via LSTM with answer position as an extra feature. The encoded representation is then fed through a gated self-matching network (Wang et al., 2017b) to aggregate information from the entire passage and embed intra-passage dependencies. Finally, a feature fusion gate (Gong and Bowman, 2018) chooses relevant information between the original and self-matching enhanced representations.

Instead of leveraging the whole context, Du and Cardie (2018) performed a pre-filtering by running a coreference resolution system on the context passage to obtain coreference clusters for both the input sentence and the answer. The co-referred sentences are then fed into a gating network, from which the outputs serve as extra features to be concatenated with the original input vectors.

5.4 Answer-unaware QG

The aforementioned models require the target answer as an input, in which the answer essentially serves as the focus of asking. However, in the case that only the input passage is given, a QG system should automatically identify question-worthy parts within the passage. This task is synonymous with content selection in traditional QG. To date, only two works (Du and Cardie, 2017; Subramanian et al., 2018) have worked in this setting. They both follow the traditional decomposition of QG into content selection and question construction but implement each task using neural networks. For content selection, Du and Cardie (2017) learn a sentence selection task to identify question-worthy sentences from the input paragraph using a neural sequence tagging model. Subramanian et al. (2018) train a neural keyphrase extractor to predict keyphrases of the passage. For question construction, they both employed the Seq2Seq model, for which the input is either the selected sentence or the input passage with keyphrases as target answer.

However, learning what aspect to ask about is quite challenging when the question requires reasoning over multiple pieces of information within the passage; *cf* the Gollum question from the introduction. Beyond retrieving question-worthy information, we believe that studying how different reasoning patterns (e.g., inductive, deductive, causal and analogical) affects the generation process will be an aspect for future study.

5.5 Technical Considerations

Common techniques of NLG have also been considered in NQG model, summarized as 3 tactics:

1. Copying Mechanism. Most NQG models (Zhou et al., 2017; Yuan et al., 2017; Wang et al., 2018; Harrison and Walker, 2018; Kumar et al., 2018a) employ the *copying mechanism* of Gülçehre et al. (2016), which directly copies relevant words from the source sentence to the question during decoding. This idea is widely accepted as it is common to refer back to phrases and entities appearing in the text when formulating factoid questions, and difficult for a RNN decoder to generate such rare words on its own.

2. Linguistic Features. Approaches also seek to leverage additional linguistic features that complements word embeddings, including word case, POS and NER tags (Zhou et al., 2017; Wang et al., 2018) as well as coreference (Harrison and Walker, 2018) and dependency information (Kumar et al., 2018a). These categorical features are vectorized and concatenated with word embeddings. The feature vectors can be either one-hot or trainable and serve as input to the encoder.

3. Policy Gradient. Optimizing for just ground-truth log likelihood ignores the many equivalent ways of asking a question. Relevant QG work (Yuan et al., 2017; Kumar et al., 2018b) have adopted policy gradient methods to add task-specific rewards (such as BLEU or ROUGE) to the original objective. This helps to diversify the questions generated, as the model learns to distribute probability mass among equivalent expressions rather than the single ground truth question.

5.6 The State of the Art

In Table 2, we summarize existing NQG models with their employed techniques and their best-reported performance on SQuAD. These methods achieve comparable results; as of this writing, Zhao et al. (2018) is the state-of-the-art.

Two points deserve mention. First, while the copying mechanism has shown marked improvements, there exist shortcomings. Kim et al. (2019) observed many invalid answer-revealing questions attributed to the use of the copying mechanism; *cf* the John Francis example in Section 5.1. They abandoned copying but still achieved a performance rivaling other systems. In parallel application areas such as machine translation, the copy mechanism has been to a large extent replaced with self-attention (Lin et al., 2017) or transformer (Vaswani et al., 2017). The future prospect of the copying mechanism requires further investigation. Second, recent approaches that employ paragraph-level contexts have shown promising results: not only boosting performance, but also constituting a step towards deep question generation, which requires reasoning over rich contexts.

6 Emerging Trends

We discuss three trends that we wish to call practitioners' attention to as NQG evolves to take the center stage in QG: Multi-task Learning, Wider Input Modalities and Deep Question Generation.

6.1 Multi-task Learning

As QG has become more mature, work has started to investigate how QG can assist in other NLP tasks, and vice versa. Some NLP tasks benefit from enriching training samples by QG to alleviate the data shortage problem. This idea has been successfully applied to semantic parsing (Guo et al., 2018a) and QA (Sachan and Xing, 2018). In the semantic parsing task that maps a natural language question to a SQL query, Guo et al. (2018a) achieved a 3% performance gain with an enlarged training set that contains pseudo-labeled $(SQL, question)$ pairs generated by a Seq2Seq QG model. In QA, Sachan and Xing (2018) employed the idea of self-training (Nigam and Ghani,

Models	Answer Encoding	Features					Performance		
		QW	PC	CP	LF	PG	BLEU-4	METEOR	ROUGE$_L$
Du et al. (2017)	not used						12.28	16.62	39.75
Duan et al. (2017)	not used	•					12.28	–	–
Zhou et al. (2017)	answer position			•	•		13.29	–	–
Yuan et al. (2017)	answer position			•		•	10.50	–	–
Wang et al. (2018)	answer position			•	•		13.86	18.38	44.37
Harrsion et al. (2018)	answer position			•	•		14.39	19.54	43.00
Kumar et al. (2018b)	not used			•	•	•	16.17	19.85	43.90
Sun et al. (2018)	answer+context position	•		•			15.64	–	–
Zhao et al. (2018)	answer position		•	•			**16.38**	**20.25**	**44.48**
Du and Cardie (2018)	answer position		•	•			15.16	19.12	–
Song et al. (2018)	separate encoder			•			13.98	18.77	42.72
Kim et al. (2019)	separate encoder						16.20	19.92	43.96

Table 2: Existing NQG models with their best-reported performance on SQuAD. Legend: **QW**: question word generation, **PC**: paragraph-level context, **CP**: copying mechanism, **LF**: linguistic features, **PG**: policy gradient.

2000) to jointly learn QA and QG. The QA and QG models are first trained on a labeled corpus. Then, the QG model is used to create more questions from an unlabeled text corpus and the QA model is used to answer these newly-created questions. The newly-generated question–answer pairs form an enlarged dataset to iteratively retrain the two models. The process is repeated while performance of both models improve.

Investigating the core aspect of QG, we say that a well-trained QG system should have the ability to: (1) find the most salient information in the passage to ask questions about, and (2) given this salient information as target answer, to generate an answer related question. Guo et al. (2018b) leveraged the first characteristic to improve text summarization by performing multi-task learning of summarization with QG, as both these two tasks require the ability to search for salient information in the passage. Duan et al. (2017) applied the second characteristic to improve QA. For an input question q and a candidate answer $\hat{a}$, they generate a question $\hat{q}$ for $\hat{a}$ by way of QG system. Since the generated question $\hat{q}$ is closely related to $\hat{a}$, the similarity between q and $\hat{q}$ helps to evaluate whether $\hat{a}$ is the correct answer.

Other works focus on jointly training to combine QG and QA. Wang et al. (2017a) simultaneously train the QG and QA models in the same Seq2Seq model by alternating input data between QA and QG examples. Tang et al. (2018) proposed a training algorithm that generalizes Generative Adversarial Network (GANs) (Goodfellow et al., 2014) under the question answering scenario. The model improves QG by incorporating an additional QA-specific loss, and improving QA performance by adding artificially generated training instances from QG. However, while joint training has shown some effectiveness, due to the mixed objectives, its performance on QG are lower than the state-of-the-art results, which leaves room for future exploration.

6.2 Wider Input Modalities

QG work now has incorporated input from knowledge bases (KBQG) and images (VQG).

Inspired by the use of SQuAD as a question benchmark, Serban et al. (2016) created a 30M large-scale dataset of *(KB triple, question)* pairs to spur KBQG work. They baselined an attention seq2seq model to generate the target factoid question. Due to KB sparsity, many entities and predicates are unseen or rarely seen at training time. ElSahar et al. (2018) address these *few-/zero-shot* issues by applying the copying mechanism and incorporating textual contexts to enrich the information for rare entities and relations. Since a single KB triple provides only limited information, KB-generated questions also *overgeneralize* — a model asks "Who was born in New York?" when given the triple *(Donald_Trump, Place_of_birth, New_York)*. To solve this, Khapra et al. (2017) enrich the input with a sequence of keywords collected from its related triples.

Visual Question Generation (VQG) is another emerging topic which aims to ask questions given an image. We categorize VQG into *grounded-* and *open-ended* VQG by the level of cognition. Grounded VQG generates *visually grounded* questions, *i.e.*, all relevant information for the answer can be found in the input image (Zhang et al., 2017). A key purpose of grounded VQG is to support the dataset construction for VQA. To ensure the questions are grounded, existing systems rely on image captions to varying degrees. Ren et al. (2015) and Zhu et al. (2016) simply convert image captions into questions using rule-based methods with textual patterns. Zhang et al. (2017) proposed a neural model that can generate questions with diverse types for a single image, using separate networks to construct dense image captions and to select question types.

In contrast to grounded QG, humans ask higher cognitive level questions about what can be inferred rather than what can be seen from an image. Motivated by this, Mostafazadeh et al. (2016) proposed open-ended VQG that aims to generate natural and engaging questions about an image. These are deep questions that require high cognition such as analyzing and creation. With significant progress in deep generative models, marked by variational auto-encoders (VAEs) and GANs, such models are also used in open-ended VQG to bring "creativity" into generated questions (Jain et al., 2017; Fan et al., 2018), showing promising results. This also brings hope to address deep QG from text, as applied in NLG: *e.g.*, SeqGAN (Yu et al., 2017) and LeakGAN (Guo et al., 2018c).

6.3 Generation of Deep Questions

Endowing a QG system with the ability to ask deep questions will help us build curious machines that can interact with humans in a better manner. However, Rus et al. (2007) pointed out that asking high-quality deep questions is difficult, even for humans. Citing the study from Graesser and Person (1994) to show that students in college asked only about 6 deep-reasoning questions per hour in a question–encouraging tutoring session. These deep questions are often about events, evaluation, opinions, syntheses or reasons, corresponding to higher-order cognitive levels.

To verify the effectiveness of existing NQG models in generating deep questions, Chen et al. (2018) conducted an empirical study that applies the attention Seq2Seq model on LearningQ, a deep-question centric dataset containing over 60% questions that require reasoning over multiple sentences or external knowledge to answer. However, the results were poor; the model achieved miniscule BLEU-4 scores of < 4 and METEOR scores of < 9, compared with > 12 (BLEU-4) and > 16 (METEOR) on SQuAD. Despite further in-depth analysis are needed to explore the reasons behind, we believe there are two plausible explanations: (1) Seq2Seq models handle long inputs ineffectively, and (2) Seq2Seq models lack the ability to reason over multiple pieces of information.

Despite still having a long way to go, some works have set out a path forward. A few early QG works attempted to solve this through building deep semantic representations of the entire text, using concept maps over keywords (Olney et al., 2012) or minimal recursion semantics (Yao and Zhang, 2010) to reason over concepts in the text. Labutov et al. (2015) proposed a crowdsourcing-based workflow that involves building an intermediate ontology for the input text, soliciting question templates through crowdsourcing, and generating deep questions based on template retrieval and ranking. Although this process is semi-automatic, it provides a practical and efficient way towards deep QG. In a separate line of work, Rothe et al. (2017) proposed a framework that simulates how people ask deep questions by treating questions as formal programs that execute on the state of the world, outputting an answer.

Based on our survey, we believe the roadmap towards deep NGQ points towards research that will (1) enhance the NGQ model with the ability to consider relationships among multiple source sentences, (2) explicitly model typical reasoning patterns, and (3) understand and simulate the mechanism behind human question asking.

7 Conclusion – What's the Outlook?

We have presented a comprehensive survey of NQG, categorizing current NQG models based on different QG-specific and common technical variations, and summarizing three emerging trends in NQG: multi-task learning, wider input modalities, and deep question generation.

What's next for NGQ? We end with future potential directions by applying past insights to current NQG models; the "unknown unknown", promising directions yet explored.

When to Ask: Besides learning what and how to ask, in many real-world applications that question plays an important role, such as automated tutoring and conversational systems, learning when to ask become an important issue. In contrast to general dialog management (Lee et al., 2010), no research has explored when machine should ask an engaging question in dialog. Modeling question asking as an interactive and dynamic process may become an interesting topic ahead.

Personalized QG: Question asking is quite personalized: people with different characters and knowledge background ask different questions. However, integrating QG with user modeling in dialog management or recommendation system has not yet been explored. Explicitly modeling user state and awareness leads us towards personalized QG, which dovetails deep, end-to-end QG with deep user modeling and pairs the dual of generation–comprehension much in the same vein as in the vision–image generation area.

References

Husam Ali, Yllias Chali, and Sadid A Hasan. 2010. Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 58–67.

Lorin W Anderson, David R Krathwohl, Peter W Airasian, Kathleen A Cruikshank, Richard E Mayer, Paul R Pintrich, James Raths, and Merlin C Wittrock. 2001. A taxonomy for learning, teaching, and assessing: A revision of blooms taxonomy of educational objectives, abridged edition. *White Plains, NY: Longman*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Benjamin Samuel Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1984. *Taxonomy of educational objectives: Handbook 1: Cognitive domain*.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2787–2795.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of bleu in machine translation research. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Yllias Chali and Sadid A. Hasan. 2015. Towards topic-to-question generation. *Computational Linguistics (CL)*, 41(1):1–20.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. In *International Conference on Web and Social Media (ICWSM)*, pages 481–490.

Wei Chen and Jack Mostow. 2009. Generating questions automatically from informational text. In *International Conference on Artificial Intelligence in Education (AIED)*, pages 17–24.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. KBQA: learning question answering over QA corpora and knowledge bases. *The Proceedings of the VLDB Endowment (PVLDB)*, 10(5):565–576.

Takshak Desai, Parag Dakle, and Dan Moldovan. 2018. Generating questions for reading comprehension using coherence relations. In *The 5th Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA@ACL)*, pages 1–10.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2067–2073.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from wikipedia. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1907–1917.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1342–1352.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 866–874.

Hady ElSahar, Christophe Gravier, and Frédérique Laforest. 2018. Zero-shot question generation from knowledge graphs for unseen predicates and entity types. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 218–228.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, Yang Liu, and Xuanjing Huang. 2018. A reinforcement learning framework for natural question generation using bi-discriminators. In *International Conference on Computational Linguistics (COLING)*, pages 1763–1774.

D Gates. 2008. Generating look-back strategy questions from expository texts. In *The Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Yichen Gong and Samuel R. Bowman. 2018. Ruminating reader: Reasoning with gated multi-hop attention. In *Workshop on Machine Reading for Question Answering@ACL*, pages 1–11.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2672–2680.

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137.

Çaglar Gülçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Daya Guo, Yibo Sun, Duyu Tang, Nan Duan, Jian Yin, Hong Chi, James Cao, Peng Chen, and Ming Zhou. 2018a. Question generation from SQL queries improves neural semantic parsing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1597–1607.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018b. Soft layer-specific multi-task summarization with entailment and question generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 687–697.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018c. Long text generation via adversarial training with leaked information. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 5141–5148.

Vrindavan Harrison and Marilyn A. Walker. 2018. Neural generation of diverse questions using answer focus, contextual and linguistic features. In *International Conference on Natural Language Generation (INLG)*, pages 296–306.

Michael Heilman. 2011. Automatic factual question generation from text. *Language Technologies Institute School of Computer Science Carnegie Mellon University*, 195.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 609–617.

Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5415–5424.

Mitesh M. Khapra, Dinesh Raghu, Sachindra Joshi, and Sathish Reddy. 2017. Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 376–385.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics (TACL)*, 6:317–328.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1106–1114.

Vishwajeet Kumar, Kireeti Boorla, Yogesh Meena, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018a. Automating reading comprehension by generating question and answer pairs. In *The Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 335–348.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018b. A framework for automatic question generation from text using deep reinforcement learning. *CoRR*, abs/1808.04961.

Igor Labutov, Sumit Basu, and Lucy Vanderwende. 2015. Deep questions without deep understanding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 889–898.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 785–794.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1203–1213.

Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee. 2010. Recent approaches to dialog management for spoken dialog systems. *Journal of Computing Science and Engineering (JCSE)*, 4(1):1–22.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

David Lindberg, Fred Popowich, John C. Nesbit, and Philip H. Winne. 2013. Generating natural language questions to support learning on-line. In *European Workshop on Natural Language Generation (ENLG)*, pages 105–114.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132.

Ming Liu, Rafael A. Calvo, and Vasile Rus. 2010. Automatic question generation for literature review writing support. In *International Conference on Intelligent Tutoring Systems (ITS)*, pages 45–54.

Ming Liu, Rafael A. Calvo, and Vasile Rus. 2012. G-asks: An intelligent automatic question generation system for academic writing support. *Dialogue and Discourse (D&D)*, 3(2):101–124.

Inderjeet Mani. 1999. *Advances in automatic text summarization*. MIT press.

Karen Mazidi and Rodney D. Nielsen. 2014. Linguistic considerations in automatic question generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 321–326.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3950–3959.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *International Conference on Information and Knowledge Management (CIKM)*, pages 86–93.

Andrew McGregor Olney, Arthur C. Graesser, and Natalie K. Person. 2012. Question generation from concept maps. *Dialogue and Discourse (D&D)*, 3(2):75–99.

Santanu Pal, Tapabrata Mondal, Partha Pakray, Dipankar Das, and Sivaji Bandyopadhyay. 2010. Qgstec system description–juqgg: A rule based approach. *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 76–79.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. Exploring models and data for image question answering. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2953–2961.

Oleg Rokhlenko and Idan Szpektor. 2013. Generating synthetic comparable questions for news articles. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 742–751.

Anselm Rothe, Brenden M. Lake, and Todd M. Gureckis. 2017. Question asking as program generation. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1046–1055.

Vasile Rus, Zhiqiang Cai, and Art Graesser. 2008. Question generation: Example of a multi-year evaluation campaign. *Online Proceedings of 1st Question Generation Workshop, NSF, Arlington, VA*.

Vasile Rus, Zhiqiang Cai, and Arthur C. Graesser. 2007. Experiments on generating questions about facts. In *Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 444–455.

Vasile Rus and James C. Lester. 2009. The 2nd workshop on question generation. In *International Conference on Artificial Intelligence in Education (AIED)*, page 808.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2010. Overview of the first question generation shared task evaluation challenge. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 45–57.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai C. Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2011. Question generation shared task and evaluation challenge - status report. In *European Workshop on Natural Language Generation (ENLG)*, pages 318–320.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai C. Lintean, Svetlana Stoyanchev, and Cristian Moldovan. 2012. A detailed account of the first question generation shared task evaluation challenge. *Dialogue and Discourse (D&D)*, 3(2):177–204.

Mrinmaya Sachan and Eric P. Xing. 2018. Self-training for jointly learning to ask and answer questions. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 629–640.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Linfeng Song, Zhiguo Wang, Wael Hamza, Yue Zhang, and Daniel Gildea. 2018. Leveraging context information for natural question generation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 569–574.

Sandeep Subramanian, Tong Wang, Xingdi Yuan, Saizheng Zhang, Adam Trischler, and Yoshua Bengio. 2018. Neural models for key phrase extraction and question generation. In *Workshop on Machine Reading for Question Answering@ACL*, pages 78–88.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3930–3939.

Duyu Tang, Nan Duan, Zhao Yan, Zhirui Zhang, Yibo Sun, Shujie Liu, Yuanhua Lv, and Ming Zhou. 2018. Learning to collaborate for question answering and asking. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1564–1574.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Workshop on Representation Learning for NLP (Rep4NLP@ACL)*, pages 191–200.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 6000–6010.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Tong Wang, Xingdi Yuan, and Adam Trischler. 2017a. A joint model for question answering and question generation. *CoRR*, abs/1706.01450.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017b. Gated self-matching networks for reading comprehension and question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–198.

Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. Qg-net: a data-driven question generation model for educational content. In *Annual ACM Conference on Learning at Scale (L@S)*, pages 7:1–7:10.

Xuchen Yao, Gosse Bouma, and Yi Zhang. 2012. Semantics-based question generation and implementation. *Dialogue and Discourse (D&D)*, 3(2):11–42.

Xuchen Yao and Yi Zhang. 2010. Question generation with minimal recursion semantics. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 68–75.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2852–2858.

Xingdi Yuan, Tong Wang, Çaglar Gülçehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *The 2nd Workshop on Representation Learning for NLP (Rep4NLP@ACL)*, pages 15–25.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. Automatic generation of grounded visual questions. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4235–4243.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3901–3910.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *CCF International Conference of Natural Language Processing and Chinese Computing (NLPCC)*, pages 662–671.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4995–5004.

This figure "Framework.png" is available in "png" format from:

http://arxiv.org/ps/1905.08949v2