Towards automatically generating Questions under Discussion to link information and discourse structure

Kordula De Kuthy **Madeeswaran Kannan**
Haemanth Santhi Ponnusamy **Detmar Meurers**

SFB 833, Project A4, University of Tübingen, Germany
{kdk,mkannan,hsp,dm}@sfs.uni-tuebingen.de

Abstract

Questions under Discussion (QUD; Roberts, 2012) are emerging as a conceptually fruitful approach to spelling out the connection between the information structure of a sentence and the nature of the discourse in which the sentence can function. To make this approach useful for analyzing authentic data, Riester, Brunetti & De Kuthy (2018) presented a discourse annotation framework based on explicit pragmatic principles for determining a QUD for every assertion in a text. De Kuthy et al. (2018) demonstrate that this supports more reliable discourse structure annotation, and Ziai and Meurers (2018) show that based on explicit questions, automatic focus annotation becomes feasible. But both approaches are based on manually specified questions.

In this paper, we present an automatic question generation approach to partially automate QUD annotation by generating all potentially relevant questions for a given sentence. While transformation rules can concisely capture the typical question formation process, a rule-based approach is not sufficiently robust for authentic data. We therefore employ the transformation rules to generate a large set of sentence-question-answer triples and train a neural question generation model on them to obtain both systematic question type coverage and robustness.

1 Introduction

As the attention in linguistics is shifting from the analysis of isolated sentences to how information is encoded in discourse, the information structure of sentences and how it is expressed in a given language is receiving increasing interest. In order to connect the information structure of sentences to the overall structure of the discourse, so-called Questions under Discussion (QUD) (Van Kuppevelt, 1995; Roberts, 2012; Beaver and Clark, 2009; Velleman and Beaver, 2016) have emerged as a way to spell out the hinge between the properties of the sentence and the nature of the discourse in which the sentence can function.

QUDs make explicit, how the meaning expressed by a sentence fits into the functional structure of the evolving discourse. Example (1) illustrates this with the implicit QUD Q_2 connecting assertion A_2 to the previous discourse A_1. Concretely, the sentence A_2 answers the question Q_2 given discourse A_1.

(1) A_1 You were working until last summer for the NSA
Q_2 *What did he do during his time at the NSA?*
A_2 and during this time you [secretly collected thousands of confidential documents.]$_F$

That part of the sentence that directly answers the QUD, here the VP *secretly collected thousands of confidential documents*, is marked as the so-called Focus (Rooth, 1992; Krifka and Musan, 2012), indicating how the alternatives implicitly present in the discourse are narrowed down as the discourse progresses. QUDs can thus be seen as a way of making alternatives explicit in the discourse.

This intuitive idea is also mentioned in corpus-based research attempting to analyze the information structure of naturally occurring data (Ritz et al., 2008; Calhoun et al., 2010). Yet, these approaches were only rewarded with limited success in terms of inter-annotator agreement, arguably because the task of

Proceedings of the 28th International Conference on Computational Linguistics, pages 5786–5798
Barcelona, Spain (Online), December 8-13, 2020

identifying QUDs was not made explicit. More recently, Ziai and Meurers (2014) and De Kuthy et al. (2016) showed that for data collected in task contexts including explicit questions, such as answers to reading comprehension questions, focus annotation becomes more reliable. The explicit question context enables experts and non-experts to reach substantial agreement in the annotation of discourse functions such as focus. In addition, automated annotation of information structure becomes feasible when explicit questions are given (Ziai and Meurers, 2018).

Bridging the gap from corpora already containing explicit questions to the analysis of any type of authentic language data, Riester et al. (2018) spell out a discourse annotation approach in which explicit pragmatic principles spell out how a QUD can be formulated for every assertion expressed by a text. De Kuthy et al. (2018; 2019) show that in corpora that are manually annotated with explicit QUDs, information structure concepts such as focus and topic can be annotated with higher inter-annotator agreement than in previous work only implicitly making use of the idea of QUDs.

While explicitly annotating corpora with QUDs appears to be a key for reliable manual or automatic annotation of information and discourse structure, in all of the above approaches it is a complex manual step. In this paper, we propose to partially automate the QUD annotation process for authentic German data. Proceeding sentence by sentence through the text, we propose to automatically generate all potentially relevant questions for a given sentence. In principle, transformation rules can transparently express the potential types of question-answer pairs, e.g., a *who* question asking for the subject of a sentence, or a *when* question asking for a temporal adverbial. But while the relationship between the question phrase and the answer phrase can concisely be expressed by such transformation rules, the selection of the proper question phrase, the identification and removal of the answer phrase, and the reformulation of the sentence into question form and word order depends on a complex interplay of factors.

We therefore proceed in two steps. We first employ a rule-based approach for German to produce questions for a large German newspaper corpus. We then pair each question with its answer phrase and the sentence it was generated from and use it to train a neural question generation model. The structure of the training data ensures high-quality question-answer congruence for a range of question types. At the same time, the variability of the authentic newspaper data that the rule-based approach was applied to provides a rich empirical basis for training a robust neural QG model that is capable of generalizing to data for which the rule-based approach fails to generate questions.

The paper is structured as follows: Section 2 provides some background on QUD annotation to establish the nature of the questions we aim to generate. Section 3 introduces the different methods we will combine in our approach, rule-based and neural QG. In section 4, we spell out how the rule-based approach is used to generate the data needed to train the neural QG, with the training being discussed in section 5. In the evaluation section 6, we illustrate the success of the methods in quantitative terms using BLEU scores by comparing the questions generated by the neural approach in relation to the rule-based method. We then show that the neural approach generates high-quality, well-formed questions for more sentences than the rule-based approach. Finally, we provide a qualitative analysis of the generated questions, also leading to ideas for future work discussed in section 7.

2 Discourse annotation with QUDs

While QUDs were originally discussed in theoretical linguistics using constructed examples, the annotation framework introduced in Riester et al. (2018) presents a method for the reconstruction of QUDs in authentic data. The characterization of a QUD for each assertion in a text is guided by explicit principles adapted from the formal semantic literature (Rooth, 1992; Schwarzschild, 1999; Büring, 2008; Büring, 2016). The three central principles defined by Riester et al. (2018) are:

Q-A-CONGRUENCE: QUDs must be answerable by the assertion(s) they immediately dominate.

Q-GIVENNESS: QUDs can only consist of given (or, at least, highly salient) material.

MAXIMIZE-Q-ANAPHORICITY: QUDs should contain as much given (or salient) material as possible.

How a QUD can be derived for a given sentence in a discourse according to these principles can be illustrated using the example (2) from an Edward Snowden interview discussed in Riester et al. (2018).

(2) A_1: Edward Snowden is in the meantime a household name for the whistleblower in the age of the internet.

Q: *#Who is Edward Snowden?* ↯Q-A-CONGRUENCE
Q: *#What happened?* ↯MAXIMIZE-Q-ANAPHORICITY
Q_2: *What did you do?*
Q: *#When were you working for the NSA?* ↯Q-GIVENNESS
Q: *#Who was working until last summer for the NSA?* ↯Q-GIVENNESS

A_2: You were working until last summer for the NSA.

Following the first sentence, A_1, we want to determine the QUD for the following sentence A_2. The only contextually appropriate QUD that can be derived for the assertion A_1 is the question Q_2. The other questions either violate Q-A-CONGRUENCE, or they are too general, conflicting with MAXIMIZE-Q-ANAPHORICITY, or they contain material not yet introduced in the context, violating Q-GIVENNESS.

The QUD annotation framework is formulated using concepts from formal pragmatics that are not language specific, so it can in principle be used to analyze data from any language. Annotation experiments based on the above principles confirm that the method can be successfully applied to German, English, and Italian data (De Kuthy et al., 2018; 2019) . The approach is shown to support information-structure annotation of authentic data with substantial inter-annotator agreement, which eliminates the road block posed by the low agreement results for information-structure annotation reported in Ritz et al. (2008).

However, manual specification of QUDs is a substantial effort making it difficult to realize large scale discourse annotation. To address this limitation and further our understanding of QUDs as a link between sentence and discourse, this paper explores the use of automatic question generation. We focus on the first step: generating all questions that can be answered by a given sentence. Selecting the question under discussion supported by the discourse from that set of questions is left to future work.

3 Automated question generation

3.1 Rule-based question generation

In computational linguistics, question generation (QG) has been tackled in several, usually applied contexts, mostly focusing on English. Automatically generating questions is a challenging task involving methods such as parsing, coreference resolution, and the transformation of syntactic structures reflecting complex linguistic characteristics. A variety of QG systems were developed, often for educational purposes, e.g., assisting students in reading (Mazidi and Nielsen, 2015), vocabulary learning (Brown et al., 2005; Mostow et al., 2004), or the assessment of reading comprehension (Le et al., 2014).

While systems for such applications are usually designed to generate a particular, task-specific set of questions, some also try to generate as many different questions as possible. Such generation traditionally involves a form of transformation, be it based on shallow rules created manually (Liu et al., 2010) or learned from data (Curto et al., 2012), syntax-based transformation rules (Heilman, 2011), or transformations based on semantic representations (Mannem et al., 2010; Yao and Zhang, 2010; Yao et al., 2012; Chali and Hasan, 2012), with some proposals also integrating discourse cues (Agarwal et al., 2011).

There is much less QG research for languages other than English, such as German. Gütl et al. (2011) focus on the extraction of concepts from German text, reporting very little on how questions are actually constructed. As far as we are aware, Kolditz (2015) is the only systematic exploration of the characteristics and challenges of QG for German. The rule-based QG system he implemented selects a potential answer phrase (NPs, PPs and embedded clauses) based on a syntactic analysis of the input sentence, replaces it with an appropriate question phrase, and transforms the syntactic representation of the declarative input sentence into question form. This is realized using a complex NLP pipeline performing constituency and dependency parsing, morphological analysis, and identification of relevant semantic characteristics. The collected information supports answer phrase selection using a set of 18 rules formulated as Tregex (Levy and Andrew, 2006) patterns. For each answer phrase thus identified, a second set of rules identifies an appropriate question word or phrase. Finally, transformation rules with a

linearization component realize the actual QG. A particular challenge is the correct insertion of the initial constituent of the input sentence into the rest of the sentence so that the initial position of the question can be occupied by the question phrase, as required by German syntax.

Failure to generate questions or incorrectly generated questions can result from a number of sources. The NLP pipeline can introduce errors in the analysis of the input sentence, or even completely fail to process it. The answer phrase selection can fail to select all potential answer phrases, the question word selection can fail to identify the correct question word for a given answer phrase, and the linearization component can fail to reorder the sentence initial constituent correctly. Kolditz (2015) evaluates the system based on a set of 93 sentences from newswire texts. The system generated questions for 77 of the 93 sentences and produced 150 questions in total, i.e., on average two per sentence. With respect to the quality of the generated sentences, manual evaluation found 60% to be well-formed. The evaluation thus confirmed that a substantial number of questions of good quality can be generated, but the approach is not sufficiently robust to produce a broad range of question types for any input sentence.

3.2 Neural question generation

Complementing rule-based QG approaches, recent QG research has focused on developing deep neural QG methods. Sequence-to-sequence (Seq2Seq) architectures (Sutskever et al., 2014), where an encoder network learns a representation of the source sequence and a decoder network generates target words, have seen significant success in sequence learning tasks such as machine translation. The inclusion of global and local attention mechanisms (Bahdanau et al., 2014; Luong et al., 2015) and self-attention (Vaswani et al., 2017), has resulted in state-of-the-art performance in several NLP tasks (Edunov et al., 2018; Raffel et al., 2019). Indeed, a recent survey of neural question generation (NQG) research (Pan et al., 2019) shows that Seq2Seq architectures also form the basis of many current QG approaches.

Du et al. (2017) conditioned their generative model on target answers by encoding the position of the answer in the context as an input feature. Where generating questions takes entire paragraphs for which a question is to be generated as input, extra information about the intent of asking reduces the ambiguity of the task. Another important challenge of QG is question word generation so that Sun et al. (2018) and Du et al. (2017) split the QG task into determining the question word or type and then generating the rest of the question. The former approach employs a template-based approach with two Seq2Seq models while the latter proposes a more flexible approach that involves learning an additional parameter during decoding that explicitly generates the question word. To handle larger contexts and deal with the problem of out-of-vocabulary words, Zhao et al. (2018) implemented a gated self-attention encoder with a Copy/Maxout pointer mechanism. Kumar et al. (2018) leveraged linguistic features such as POS and NER tags and deep reinforcement learning techniques such as policy gradient methods to add additional task-specific rewards to the training objective.

4 Obtaining data for a neural question generation approach

The productivity and rich compositionality of human language makes handcrafting rule-based systems for robust question generation very difficult. Deep neural QG methods, on the other hand, can learn latent representations of syntactic and semantic language characteristics from large sets of data. For this one needs pairs of questions and the text they are asking about, i.e., a lot of instances of the generalizations encoded by the transformation-rules in a large, variable set of authentic language data. An NQG model trained on such a rich set of data then can potentially be more robust since it learns the systematic transformation patterns mixed in with the variable language patterns characterizing authentic use. Obtaining appropriate training corpora thus is an essential ingredient of NQG.

In current research approaching QG in the context of question answering (QA), QA corpora such as SQuAD (Rajpurkar et al., 2016), Coqa (Reddy et al., 2019), or Quac (Choi et al., 2018) are typically used to train and evaluate the NQG models. However, such corpora are not well-suited given our goal of generating QUDs. These corpora provide a paragraph-level context for each question, where the question is related to the information encoded in the paragraph, not to the way this information is structured and presented in a sentence. So Q-A-Congruence between the question and a sentence that answers it is not

ensured. For research like ours that investigates the relationship between the information structure of a sentence and the question context, corpus data that does not ensure Q-A-Congruence is insufficient.

In addition, the existing corpora overwhelmingly provide only English data. The few multilingual, parallel datasets such as XQUAD (Artetxe et al., 2019) and MLQA (Lewis et al., 2019) are evaluation datasets of very limited size. Automatically translating a corpus or designing a neural model architecture to jointly translate, align and generate questions (Carrino et al., 2019), while potentially promising, substantially increases complexity and potentially reduces performance due to translation error propagation.

As empirical basis for developing a German NQG approach with the vision of generating QUDs, we created a suitable corpus by leveraging the rule-based question generation system of Kolditz (2015), which he kindly made available to us. Creating such a corpus required a large, authentic German text source. We initially considered using the German version of Wikipedia. Yet the encyclopedic nature of these texts collecting facts from a wide of perspectives makes them suboptimal for our approach designed to identify QUDs in a text with a coherent discourse structure. We therefore settled on the German newspaper *Die Tageszeitung* (TAZ, https://taz.de) which in the science edition is available in XML format and has also been used for the German TüBa-D/Z treebank (Telljohann et al., 2004).

We extracted the text of 450K individual TAZ articles from years 1995 to 2001 using *Beautiful Soup 4* (https://crummy.com/software/BeautifulSoup) and performed tokenization and sentence segmentation using *spaCy*'s (https://spacy.io) de_core_news_sm model. To eliminate potential segmentation problems and focus on declarative sentences, we filtered out sentences with fewer than four tokens, not starting with an uppercase letter, or not ending with a period or exclamation mark. The resulting 5.46 million sentences were fed into an updated version of the rule-based QG of Kolditz (2015), producing a corpus of 5.24 million triples of the form <sentence, question, and the answer phrase in the sentence given the question> (which we make available upon request). This includes over 30 different types of question phrases, also discussed in section 6. The most common types of answer phrases are NP subjects and objects as well as various types of adverbial modifiers. (3) and (4) show two typical examples.

(3) A: **Beamte, Richter und Soldaten in Ostdeutschland** werden auch in Zukunft weniger
civil servants, judges and soldiers in East Germany will also in future less
verdienen als ihre westdeutschen Kollegen.
earn than their West German colleagues
Civil servants, judges and soldiers in East Germany will continue to earn less than their West German colleagues.

Q: **Wer** wird auch in Zukunft weniger verdienen als ihre westdeutschen Kollegen?
who will also in future less earn than their West German colleagues
Who will continue to earn less than their West German colleagues in the future?

(4) A: Nostalgiker erzählen noch gerne **von dem "Silbertäßchen" Medellin.**
nostalgics talk still like about the little silver cup Medellin
Nostalgics still like to tell about the little silver cup of Medellin.

Q: **Wovon** erzählen Nostalgiker noch gerne?
what of talk nostalgics still like
What else do nostalgics like to tell about?

In (3), the complex subject NP shown in bold is replaced by the matching question word *Wer* ('who'), and the question is generated with adjusted agreement morphology on the finite verb (*werde* → *wird*). In (4), the PP object in bold is replaced by the question word *Wovon* ('what of'), and the question appropriately integrates the originally sentence-initial phrase (*Nostalgiker*).

5 Training a neural question generation model

Question generation is a sequence learning problem where the model accepts an input sequence $x_1, ..., x_n$ and learns the conditional probability $p(y|x, z)$ of generating the target question $y_1, ..., y_m$ while conditioned on the answer z:

$$log\ p(y \mid x, z) = \sum_{j=1}^{m} log\ p(y_j \mid y_{<j}, x, z)$$

We implemented a Seq2Seq model with multiplicative attention (Luong et al., 2015) using *TensorFlow* 2.0 (Abadi et al., 2015), with our code available upon request. The (surface-form) tokens of the source sentence, their part-of-speech tags, and the span of the answer phrase were used as inputs to the model. *spaCy* (https://spacy.io) with the de_core_news_sm pretrained model was used for tokenization, tagging, and parsing. The answer span was encoded in IOB format. All input sequences were padded with special leading and trailing tokens to indicate their beginning and end. In the encoder stage of the model, the input at each timestep was the concatenation of the embeddings of the token and the POS tag, and the answer span indicator. Pretrained *fastText* embeddings (Bojanowski et al., 2017) were used to initialize the token embedding matrix, which was then frozen during training. The embedding matrix for the POS tags was randomly initialized. A fixed vocabulary was used for both input and target sequences, which is generated from 100K most frequent words in the corpus. Out-of-vocabulary (OOV) tokens were replaced with a special marker token. The model hyperparameters we used are given in Table 3 in the appendix.

We observed that many generated questions could be considered well-formed were it not for the appearance of the special token in place of OOV words. Most such occurrences appear in noun phrases, often involving proper nouns. To remedy this, we developed a simple post-processing module to heuristically improve the model's prediction. Whenever the OOV marker token appears in the generated question, the post-processing locally aligns the dependency parse tree of the noun phrases at the level of triplets (Head, Label, Modifier) between the source sentence and the generated question. The OOV marker tokens in the question then are replaced with the aligned original tokens in the input sentence. Given the heuristic nature of this align-and-copy post-processing step, its performance is limited by its local scope, the accuracy of the dependency parse, and the quality of the generated question.

The corpus introduced in section 4 was iteratively undersampled to create multiple sets of training, validation and test data for different sample sizes with the same distribution of question types. We trained two versions of the model, one on 150K training samples and the other on 400K. Validation datasets of 15K samples were used for both models. Teacher forcing was enabled to ensure training stability.

6 Evaluation

6.1 Quantitative BLEU score evaluation on large test set

In the first evaluation, the trained models predicted the questions for 14.700 previously-unseen sentences and the results were compared to gold-standard questions. For this evaluation on a large test set, for which no manually validated gold-standard questions are available, we used the questions generated by the rule-based approach as the gold standard. As measure we used the BLEU metric (Papineni et al., 2002) standardly employed in current QG research. The *SacreBLEU* (Post, 2018) Python library (V. 1.4.10 with default parameters) was used to calculate the cumulative and individual *n*-gram precision scores. Table 1 lists the results for models differing in training set size as well as the *No Copy* performance of the models without OOV post-processing step.

Model Variant	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU (Cumulative)
150K *No Copy*	79.8	66.4	56.2	47.9	61.47
150K	87.8	76.3	67.7	60.7	72.46
400K *No Copy*	84.9	75.0	67.1	60.3	71.25
400K	93.8	86.5	81.0	76.5	**84.24**

Table 1: Evaluating the Seq2Seq models in relation to the questions generated by the rule-based approach

A high BLEU score here indicates a high similarity between the question predicted by the neural model and the questions generated by the rule-based approach. The very high scores show that the model learns that QG in the QUD context can essentially be decomposed into the tasks of question phrase generation and reordering of words from the input sentence. Some generation is also needed for verb forms, as was illustrated by (3). The post-processing step successfully takes care of many OOV cases.

When put next to the results of recent NQG models for English, such as the BLEU-4 scores of 16.20 (Kim et al., 2019), 16.38 (Zhao et al., 2018), and 22.17 (Chan and Fan, 2019), the scores we presented above compare very favorably. While the scores provide some related context, they are not directly comparable in that we use a different, German corpus and base question generation on individual sentences, whereas the English approaches uses larger paragraph-sized contexts, conceptualizing QG more as a subtask of reading comprehension and question answering.

6.2 Quantitative manual evaluation on 500 sentence test set

To determine whether the high BLEU scores achieved by our models when compared to the questions generated by the rule-based method actually indicates the models' ability to generate high quality, well-formed questions, we manually evaluated the questions generated by the different neural models on a test set of 500 sentences randomly sampled from the original TAZ corpus introduced in section 4. We also ran the rule-based QG approach on this test set to compare its performance to that of the neural models.

For the 500 sentence test corpus, the rule-based system successfully generated questions for 290 sentences. To run the Seq2Seq QG models, we took the 290 sentences with the answer phrases selected by the rule-based approach and for the other 210 sentences manually selected one answer phrase each. For the resulting set of 500 sentences with marked answer phrases, each Seq2Seq model successfully generated 500 questions.

To manually evaluate the quality of the questions, we formulated annotation guidelines spelling out the aspects to be considered for the binary decision. To be judged positively, the question must be grammatically well-formed and meaningful, the sentence must answer the question (Q-A-Congruence), and the generated question phrase must match the marked answer phrase. Evaluating the inter-annotator agreement of two annotators for 200 cases, we found substantial agreement (κ = 0.74). Disagreements arose around the relevance of capitalization and punctuation, for which the guidelines were extended.

On this basis, the human annotators evaluated the questions generated by the different approaches for the 500 test sentences, with the results being shown in Table 2.

	Rule-based	150K (*No Copy*)	150K	400K (*No Copy*)	400K
Number of questions	290	500	500	500	500
Well-formed questions	183	147	223	157	**261**

Table 2: Evaluation results on random sample of 500 sentences

Of the 290 questions that the rule-based approaches generated, 183 were found to be high-quality. So when the rule-based approach generates a question, 63% of those are of high quality. This arguably supports use of those questions as reference for the BLEU evaluation in section 6.1. Overall, for the 500 sentence test set, the rule-based approach only generated 183 well-formed questions though (37%).

Of the 500 questions generated by the best neural question model (400K), 261 were found to be high quality. With 52%, the approach substantially outperforms the rule-based approach, with the success clearly being due to the robustness of the neural approach, generating questions for all sentences.

6.3 Qualitative analysis

Turning to a more in-depth investigation of the generated questions, we found that both Seq2Seq models produced questions that are identical to those of the rule-based approach, as illustrated by (5).

(5) A: Auch Otto Graf Lambsdorf ist **gegen zweierlei Wahlrecht.**
also Otto Graf Lambsdorf is against double voting rights
Otto Graf Lambsdorf is also against double voting rights.

Q: **Wogegen** ist auch Otto Graf Lambsdorf?
what against is also Otto Graf Lambsdorf
What is also Otto Graf Lambsdorf against?

For sentence A and the given answer phrase (shown in bold), both Seq2Seq models produced the question Q, with appropriate question phrase *Wogegen* ('against what') and the sentence initial phrase *auch Otto Graf Lambsdorf* ('also Otto Graf Lambsdorf') properly integrated into the question.

For the vision of generating possible QUDs, an approach should be able to generate a broad variety of different question types reflecting the possible answer phrases in a sentence, i.e., possible foci. Question type variety can be most easily measured in terms of the variety of question phrases generated by the systems. We found that the rule-based system generated high quality questions with 20 different question phrases, the 150K model 25 different ones, and the 400K model with 27 produced the largest variety. A complete table of all question phrases generated by the systems can be found in Table 4 in the appendix.

Turning from the variety to the quality of the question phrase, (6) shows a case, where both neural models produced an appropriate question word (NQ:), whereas the rule-based approach did not (RQ:).

(6) A: Bürgermeister Karl Henselin hielt sich **an die Absprache.**
Mayor Karl Henselin kept to the agreement.

NQ: **Woran** hielt sich Bürgermeister Karl Henselin?
to what kept himself mayor Karl Henselin?
What did mayor Karl Henselin keep to?

RQ: **Wohin** hielt sich Bürgermeister Karl Henselin?
where to kept himself mayor Karl Henselin?

Predicting the appropriate question phrase matching the given answer phrase in the sentence is a challenge for all QG approaches. For the rule-based model, errors in parsing the input sentence and word order errors are the two other major problems. For the neural models, unknown words, inappropriately generated semantically related words, and word order errors are the major challenges. A complete list of the error types we identified in the 500 sample questions can be found in Table 5 in the appendix.

While word order errors are among the major error sources, the Seq2Seq models in principle picked up the general word order patterns of questions in German and were even able to produce the correct question word order for cases where the rule-based system did not. This is illustrated by example (7).

(7) A: Die für heute vorgesehene Wahl der Stadtregierung "fällt erst mal flach", erklärte **ein**
the for today planned election of the city government falls for now flat announced a
CDU-Sprecher.
CDU spokesperson
The election of the city government planned for today "falls flat for now", announced a CDU spokesperson.

NQ: **Wer** erklärte die für heute vorgesehene Wahl der Stadtregierung "fällt erst mal flach"?
who announced the for today planned election of the city government falls for now flat
Who announced the election of the city government planned for today "will fall flat for now"?

RQ: **Wer** für heute vorgesehene Wahl der Stadtregierung "fällt erst mal flach", erklärte die?
who for today planned election of the city government falls for now flat announced the

In the question NQ, the neural model correctly reordered the entire subordinate clause *Die für heute vorgesehene Wahl der Stadtregierung "fällt erstmal flach"* after the finite verb *erklärte* ('announced'). In the rule-based question RQ, only the first word *Die* of this clause was reordered and occurs after the finite verb, resulting in an ill-formed question.

Since the seq2seq models are based on word embeddings, there is one error type that the rule-based system does not produce, namely the occurrence of words in the question than are somewhat but not sufficiently semantically related to words in the source sentence. Example (8) illustrates such a case.

(8) A: Die Landwirte haben auf rund 24 Hektar Schlafmohn angebaut.
the farmers have on about 24 hectares opium poppies cultivated
The farmers have cultivated opium poppies on about 24 hectares.

NQ: Wer hat auf rund 24 Hektar Schlafmohn exportiert?
who has on around 24 hectares opium poppies exported?

In the question NQ, the main verb *exportiert* ('exported') was generated by the neural question generator instead of the verb *angebaut* ('cultivated') of the input sentence. The result is an inappropriate shift in meaning that violates Q-A-Congruence.

Finally, for very long source sentences (> 20 tokens), some of the typical problems of neural generation models occur in the generated questions, such as degenerate sequences with repeated words. But overall, the neural Seq2Seq question generation models do very well with respect to the core components of the task: They learn to predict the correct question phrases for the specified answer phrase, and they pick up on the underlying German language characteristics, especially regarding the word order, that enable them to generate well-formed German questions from the authentic, highly variable input sentences.

7 Conclusion and Outlook

Based on the insight that texts annotated with QUDs are an important source for reliable discourse analysis of larger corpora, we set out to explore how the labor-intensive manual annotation of QUDs can be partially automated by generating questions for every sentence in a given text. We employed a rule-based question generation model to generate a large corpus of sentence-question-answer triples. The corpus was used to train and test a neural question generation model that, given a sentence and a possible answer phrase, generates the matching question. The quantitative analysis using BLEU scores and a manual evaluation of a sample set of generated questions showed that the neural model successfully generated meaningful, well-formed questions. It learned to predict correct question words for a given answer phrase and generated questions reflecting the appropriate word order characteristics.

As a next step, we envision a question generation model that is able generate all possible questions that can be answered by a single sentence. For fully-automated QUD annotation, the remaining step then is the selection of the specific question that is the appropriate QUD in a given discourse. As an alternative, we also plan to investigate whether the neural question generation approach can be extended with discourse context to support QUD generation in a single step. In terms of future work on the neural network approach, we also plan to leverage the insight that the vocabulary of the generated questions almost entirely derives from that of the source sentence by explicitly encoding it in the neural network's architecture using pointer networks (Vinyals et al., 2015) or copy mechanisms (Gulcehre et al., 2016). Using pretraining methods (Devlin et al., 2018) to impart the model with a deeper understanding of the problem domain is another potential avenue for this future research.

Developing an approach supporting the generation of a specific QUD for each sentence in a given discourse can also be seen as a step towards a normal form realization of QUDs. Such normal forms could facilitate automatic comparison of discourse annotations, which currently is hampered by the fact that the pragmatic principles formulated by De Kuthy et al. (2019) characterize the meaning to be encoded by the question, so often there are multiple options for the concrete formulation of the question.

Acknowledgements

We are grateful to Tobias Kolditz for making his QG system available and to Lukas Stein for his help in evaluating our systems. The work in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — SFB 833 — Project ID 75650358. We also acknowledge support through BMBF project W143500 and from the Cluster of Excellence "Machine Learning — New Perspectives for Science", EXC 2064/1, project number 390727645.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas,

Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Manish Agarwal, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Portland, OR. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

David I Beaver and Brady Z Clark. 2009. *Sense and sensitivity: How focus determines meaning*, volume 12. John Wiley & Sons.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Ann Arbor, MI. Association for Computational Linguistics.

Daniel Büring. 2008. What's new (and what's given) in the theory of focus? In *Proceedings of the 34th Annual Meeting of the Berkeley Linguistics Society*, pages 403–424.

Daniel Büring. 2016. *Intonation and Meaning*. Oxford University Press.

Sasha Calhoun, Jean Carletta, Jason Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44:387–419.

Casimiro Pio Carrino, Marta R Costa-jussà, and José AR Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering. *arXiv preprint arXiv:1912.05200*.

Yllias Chali and Sadid A Hasan. 2012. Towards automatic topical question generation. In *COLING*, pages 475–492, Mumbia, India.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent bert-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Sérgio Curto, Ana Cristina Mendes, and Luísa Coheur. 2012. Question generation based on lexico-syntactic patterns learned from the web. *Dialogue & Discourse*, 3(2):147–175.

Kordula De Kuthy, Ramon Ziai, and Detmar Meurers. 2016. Focus annotation of task-based data: Establishing the quality of crowd annotation. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 110–119, Berlin, Germany. ACL.

Kordula De Kuthy, Nils Reiter, and Arndt Riester. 2018. Qud-based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, JP.

Kordula De Kuthy, Lisa Brunetti, and Marta Berardi. 2019. Annotating information structure in Italian: Characteristics and cross-linguistic applicability of a QUD-based approach. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 113–123, Florence, Italy, August. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. *arXiv preprint arXiv:1603.08148*.

Christian Gütl, Klaus Lankmayr, Joachim Weinhofer, and Margit Hofler. 2011. Enhanced automatic question creator–eaqc: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, 9(1):23–38.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6602–6609.

Tobias Kolditz. 2015. Generating questions for German text. Master thesis in computational linguistics, Department of Linguistics, University of Tübingen.

Manfred Krifka and Renate Musan, editors. 2012. *The Expression of Information Structure*, volume 5 of *The Expression of Cognitive Categories*. De Gruyter Mouton, Berlin/Boston.

Vishwajeet Kumar, Ganesh Ramakrishnan, and Yuan-Fang Li. 2018. A framework for automatic question generation from text using deep reinforcement learning. *arXiv preprint arXiv:1808.04961*.

Nguyen-Thinh Le, Tomoko Kojiri, and Niels Pinkwart. 2014. Automatic question generation for educational applications – the state of art. In Tien van Do, Hoai An Le Thi, and Ngoc Thanh Nguyen, editors, *Advanced Computational Methods for Knowledge Engineering*, volume 282 of *Advances in Intelligent Systems and Computing*, pages 325–338. Springer International Publishing, Cham, Switzerland.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

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

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. 2010. Question generation from paragraphs at upenn: Qgstec system description. In *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91, Pittsburgh, PA.

Karen Mazidi and RodneyD. Nielsen. 2015. Leveraging multiple views of text for automatic question generation. In Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo, editors, *Artificial Intelligence in Education*, volume 9112 of *Lecture Notes in Computer Science*, pages 257–266. Springer International Publishing, Cham, Switzerland.

J. Mostow, J. E. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri. 2004. Using automated questions to assess readingcomprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning*, 2(1-2):97–134.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *arXiv preprint arXiv:1905.08949*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Arndt Riester, Lisa Brunetti, and Kordula De Kuthy. 2018. Annotation guidelines for questions under discussion and information structure. In Evangelia Adamou, Katharina Haude, and Martine Vanhove, editors, *Information structure in lesser-described languages: Studies in prosody and syntax*, Studies in Language Companion Series. John Benjamins.

Julia Ritz, Stefanie Dipper, and Michael Götze. 2008. Annotation of information structure: An evaluation across different types of texts. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2137–2142, Marrakech, Morocco.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69, December.

Mats Rooth. 1992. A theory of focus interpretation. *Natural Language Semantics*, 1(1):75–116.

Roger Schwarzschild. 1999. GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, 7(2):141–177.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lissabon.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(01):109–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Leah Velleman and David Beaver. 2016. Question-based models of information structure. In Caroline Féry and Shinichiro Ishihara, editors, *The Oxford Handbook of Information Structure*, pages 86–107. Oxford University Press.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

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, Pittsburgh, PA. Citeseer.

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

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

Ramon Ziai and Detmar Meurers. 2014. Focus annotation in reading comprehension data. In *Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII, 2014)*, pages 159–168, Dublin, Ireland. COLING, ACL.

Ramon Ziai and Detmar Meurers. 2018. Automatic focus annotation: Bringing formal pragmatics alive in analyzing the Information Structure of authentic data. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Langua ge Technologies (NAACL-HLT)*, pages 117–128, New Orleans, LA. ACL.

Appendices

Hyperparameter	Value
Batch size	128
Epochs (w/t early-stopping)	40
RNN Unit	LSTM
Encoder/Decoder Hidden Size	512
Encoder/Decoder Dropout	0.5
Word Embedding Dim	300
POS Embedding Dim	100
Beam Width	5

Table 3: Seq2Seq Model Hyperparameters

Question Phrase	400K	150K	R-based
Was (‘what’)	96	91	67
Wer (‘who’)	90	69	61
Wann (‘when’)	11	11	8
Worin (‘where in’)	9	7	8
Wo (‘where’)	7	8	7
Wozu (‘what for’)	6	6	5
Wovon (‘of what’)	5	6	3
Wen (‘whom$_{acc}$’)	4	2	4
Für wen (‘for whom’)	3	0	1
Wem (‘whom$_{dat}$’)	3	4	3
Wobei (‘where by’)	3	3	3
Wohin (‘where to’)	3	3	4
Womit (‘with what’)	3	2	2
Mit wem (‘with whom’)	2	0	0

Question Phrase	400K	150K	R-based
Warum (‘why’)	2	1	0
Wonach (‘after what’)	2	2	2
Woran (‘on what’)	2	2	0
Bei wem (‘by whom’)	1	1	1
Laut was (‘according to what’)	1	1	1
Seit wann (‘since when’)	1	2	1
Um was (‘what about’)	1	1	0
Von wem (‘of whom’)	1	0	0
Wie lange (‘how long’)	0	1	0
Wofür (‘what for’)	1	1	1
Wogegen (‘against what’)	1	1	1
Worauf (‘on what’)	1	1	0
Worüber (‘about what’)	1	1	1
Zu wem (‘to whom’)	1	1	0

Table 4: Types and Frequency of Question Phrases in Well-formed Questions of the 500 Sample

Error Type	400K	150K	R-based
W-Word	88	82	34
Unknown Word	47	52	0
Word Order	40	33	27
Answer Phrase	18	21	29
Different Word	18	58	0
Missing Word	6	7	0
Verb Form	6	5	1
Repeated Word	5	12	0
Additional Word	4	0	0
No Clause	4	4	6
Collocation	3	3	3

Table 5: Most Frequent Error Types in Questions of the 500 Sample