Conversational Question Generation in Russian

Olesia Makhnytkina, Anton Matveev, Aleksei Svischev, Polina Korobova,
Dmitrii Zubok, Nikita Mamaev, Artem Tchirkovskii
ITMO University
St. Petersburg, Russia
{makhnytkina, aymatveev, ansvishchev, pikorobova, zubok, nmamaev, chirkovskii}@itmo.ru

Abstract—**In this paper, we discuss possibilities of automatic generation of conversational questions in Russian. We are exploring the possibility of using "A Conversational Question Answering Challenge" (CoQA) dataset translated into Russian for training an encoder-decoder model. We review several techniques for improving the quality of questions generated in the Russian language. The results are evaluated manually. Combining a neural network-based approach with a rules-based approach, we develop a system for automatic examination of university students.**

I. Introduction

Dialogue systems are being actively developed and integrated into various areas of human activity, including education. The rapid increase in the number of students stimulates the adoption of remote learning and requires higher quality and variety of available material [1]. One of the most important procedures in online learning is testing. Typically, testing is performed with quizzes measuring theoretical expertise, crafting of which may be partly automated (see automatic question generation). Automating a conversational way of testing the learner's knowledge is less explored and more complicated in theory. New approaches aimed at generating questions in a conversation require large amounts of training data, which is a massive obstacle, especially when considering languages with less textual data available, such as Russian. At the moment of writing, our research found no works on conversational automatic question generation (CQG) in Russian. CQG and automatic question generation, in general, have a variety of knowledge acquisition-related functions, such as focusing learners' attention on key points, as well as repeating core ideas.

II. Related work

Currently, several approaches have been explored for generating questions automatically. Earlier works on automatic question generation rely on creating linguistic patterns and templates [2], [3]. This requires studying the linguistic traits of the text, as well as keeping track of the terms, synonyms, and hyponyms for creating topic-specific rules. This approach also requires significant time and effort to adapt to a new domain or language.

For term extraction, researchers have adopted both machine learning-based and rule-based methods. For example, there are methods based on lexico-syntactic templates [4]–[6], Ripper's algorithm for term extraction [7], support vector machine combined with morphological rules [8], [9], and convolutional neural networks [10].

Few works have been published on hyponym extraction. The method by Luan employs lexico-syntactic structures (LSPE) [11]. In work, authors present an example of hyponym extraction from The New York Times newspaper and describe a method for automated detection of lexico-semantic relations WordNetstyle by parsing large text datasets with lexico-semantic templates. For creating a knowledge base covering a branch of humanities, one could perform named entity recognition as well as extracting dates and time spans; the KERA algorithm is capable of that [8].

A study [12] aims to automatically acquire a col-lection of hypernym-hyponym word pairs in Bahasa Indone-sia using pattern analysis approach. The Wikipedia text isutilized as the data source to obtain the hypernym-hyponymrelations. The presented method to acquire relations isdefined in systematic procedures, which are seed building,text pre-processing, sentence extraction, pattern extraction,and pattern matching. The paper [13] describes a rule-based approach for hypernym and hyponym extraction from Russian texts. For this task was employed finite state transducers (FSTs), was developed 6 finite state transducers that encode 6 lexicosyntactic patterns, which show a good precision on Russian DBpedia: 79.5 percent of the matched contexts are correct.

Conversely, modern approaches to automatic question generation leverage deep learning models, as per review by Kurdi et al. [14]; most frequently, recurrent neural networks [15] and specifically LSTM [16]. The work by Xiao et al. explores domain adaptation of such models with domain-specific dictionaries [17]. Both rule-based and machine learning-based methods have their advantages. Rule-based systems are more predictable and allow full control of dialogue flow, while ML-based systems can more easily adapt to new knowledge domains and require less linguistic 'programming'. As such, combining the two approaches makes sense for a more robust solution. Currently, most studies on generating questions from text operate with Chinese [18] or English [19]–[21] data. Research for Russian has been mostly inactive, with the main reason being the lack of available data [22].

There is data available for performing question answering, however; such as the dataset prepared for Task B of Sberbank Data Science Journey 2017 (**SDSJ-17**). The dataset consists of question-answer pairs, each of which also has a 'context' - a passage of text the question references. The answer is also presented as a span from the text. **SDSJ-17** includes factoid questions and much fewer closed questions. For tokenization, we used the traditional word-by-word tokenization, as well as the Byte-Pair Encoding Tokenizer (https://opennmt.net/OpenNMT/tools/tokenization/#bpe) which allows learning embeddings for sub-word tokens, providing the benefit of retrieving embeddings for new words. We try several approaches to text generation: beam search with copying, top-k sampling, and temperature sampling.

In English, there are two datasets available both of which are larger: SQuAD dataset created by crowd workers using Wikipedia pages [15]; and MS MARCO, formed using Microsoft's search engine Bing [23].

Here are some examples of work on automatic question generation. Killawala et al. [24] attempted to use an LSTM network, named entity recognition, and Super-sense tagging to construct true/false, gap-fill multiple-choice, and factual questions. Du et al. [25] and Zhao et al. [26] use a seq2seq architecture with copying to generate factual questions in an end-to-end approach. Scialom et al. [27] also implement copying for a Transformer network. While AQG is not their goal but rather an approach to training a QA system, Duan et al. [28] describe a four-step method of generating questions in both retrieval and generative manner. Cho et al. [29] additionally feed POS and NER data into an encoder-decoder network, along with introducing a so-called selector module, which enables diversification during learning. Along with other NLP tasks, Dong et al. [30] evaluate a Transformer jointly pretrained on multiple language modeling objectives on AQG, achieving a high BLEU-4 score on SQuAD 1.1. Yan et al. [31] introduce an n-gram prediction objective for pre-training a Transformer-based language model, reaching state-of-the-art on SQuAD 1.1 at the time of writing.

CQG is a novel task, which gained interest with the newly released CoQA dataset [32]. Gao et al. [33] encode the passage and conversation history jointly; they also investigate conversation flow and the effect of coreference alignment. Pan et al. [34] use cross-attention and employ policy gradient with a separate QA model, which contributes to a higher BLEU score for the output questions.

Inspired by the many efforts and insights on online learning provided in recent years, we attempt to design a system for automatic examination of university students. Our system introduces a virtual agent, whose purpose is to assess the level of students' knowledge by asking questions and evaluating answers in a dialogue. The work by Matveev et al. [35] presented results of implementing such a system.

Here, we discuss the potential of performing automatic conversational question generation in Russian. More specifically, the task we consider is generating an appropriate question that can be answered after reading the provided text. The task was first formulated in the work of Pan et al., which proposes an architecture for generating questions in English [36]; we intend to build upon some of their ideas.

III. METHODS & DATA

A. Data

As the source of data, we use text materials on the subject of Natural Language Processing. The texts contain terms as well as their definitions and synonyms, hyponym-hypernym relations, which were leveraged for creating a knowledge structure to help question generation by using handcrafted templates. By "term" we consider a word or phrase which in context describes a well-defined scientific concept [37], [38]; as its synonym we consider a word or phrase used alternatively to describe the said concept. By hyponym, we mean a term of more specific meaning than a general or superordinate term applicable to it, e.g.: 'Some of the formal languages are the language of mathematical logic, programming languages, languages derived from regular expressions, etc.'. Here the terms 'language of mathematical logic', 'programming languages', 'languages derived from regular expressions' are hyponyms of the term 'formal languages'.

Such markup (see table I) also allows us to evaluate the quality of relations extracted for creating the knowledge base. By having labels for each sentence describing the ground truth relations, we could count the True Positives (the number of predicted relations for a sentence which are correct), False Negatives (the number of correct relations that were not predicted), and False Positives (the number of predicted relations not found in the label).

TABLE I. EXAMPLE OF DATA MARKUP

Text	**Markup**
Natural language (NL) is a language that people use to talk which was not created artificially. Examples of natural languages are Russian, English, Chinese, Kazakh, etc. Languages used by people to communicate are often contrasted with formal languages. Examples of formal languages: mathematical logic; programming languages; languages, generated by regular expressions, finite-state machines, Chomsky hierarchy, etc. Artificially created for a certain purpose languages are called "constructed languages". There are over 1000 such languages and growing. Some examples: Esperanto, Lojban, Toki Pona, Elf, etc.	TERM(natural language), HYPERNYM(language), SYNONYM(NL) TERM(natural language), EXAMPLE(Russian), EXAMPLE(English), EXAMPLE(Chinese), EXAMPLE(Kazakh), TERM(languages), TERM(formal language). TERM(formal language), EXAMPLE(mathematical logic), EXAMPLE(programming language), TERM(regular expression), TERM(finite-state machine), TERM() TERM(language), HYPONYM(constructed language) TERM(constructed language), EXAMPLE(esperanto), EXAMPLE(lojban), EXAMPLE(toki pona), EXAMPLE(elf)

For generating questions in a conversation, i.e. considering both the user's answer and the conversation context, we used CoQA dataset, which contains 127k questions with answers from more than 8k conversations. Each conversation was performed by a pair of crowd workers discussing a text passage; as such, most questions are 'conversational', meaning they use the knowledge gained from the previous conversation. Answers are free-form; each answer is paired with a rationale or a span of the passage supporting it.

The dataset is split into conversations, each including a passage of text, as well as a series of questions and answers which are rationales, or spans pulled from the text. **CoQA** mostly includes either closed questions (yes/no questions) or factoid questions, the answers to which are expressions referring to a person, object, time, or location. The dataset was translated to Russian using an online translation service, Yandex.Translate (Fig. 1). Questions without an answer were dropped. We also expanded some insufficient rationales.

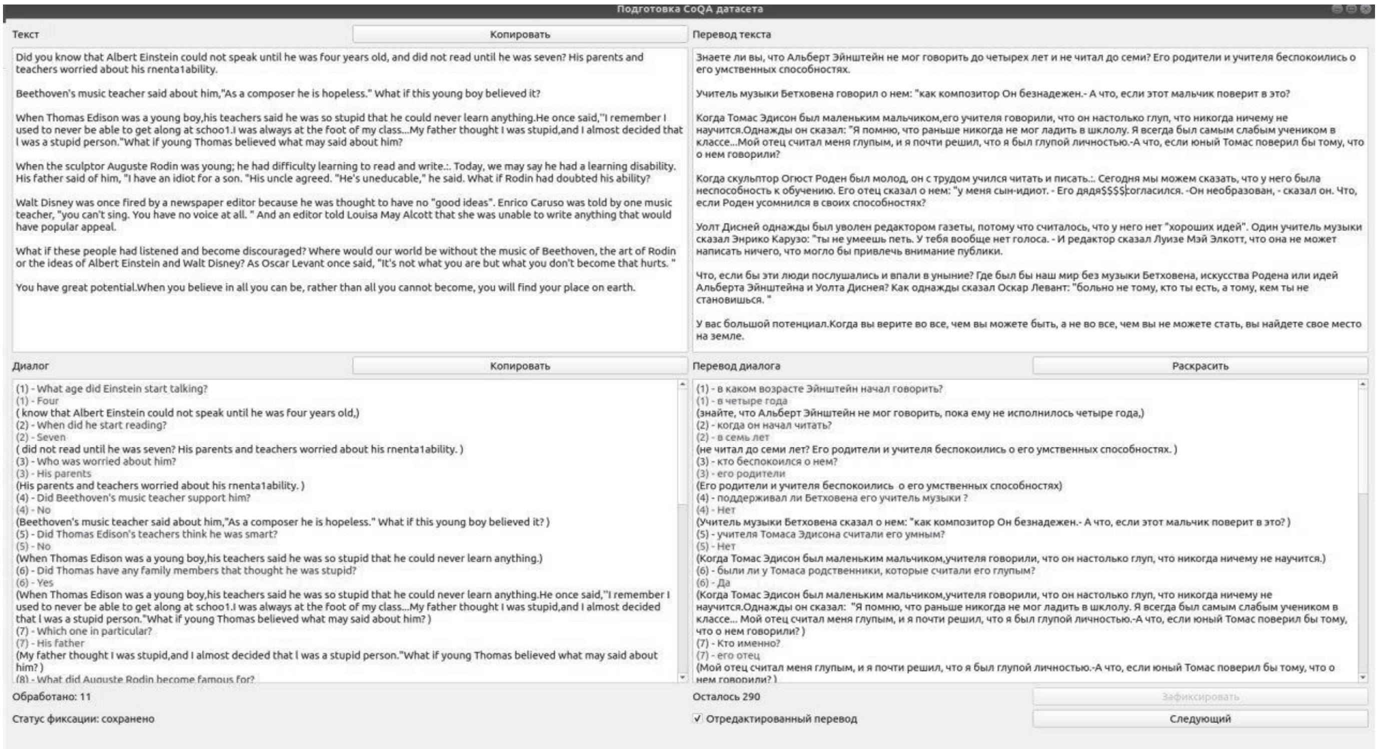
The dataset translated to Russian was also augmented with dialogues simulating an oral exam:

Examination question:
Natural language processing **A fragment from lecture notes with the answer:**
Natural language (NL) is a language that people use to talk which was not created artificially. Examples of natural languages are Russian, English, Chinese, Kazakh, etc. Languages used by people to communicate are often contrasted with formal languages. Examples of formal languages: mathematical logic; programming languages; languages, generated by regular expressions, finite-state machines, Chomsky hierarchy, etc. Artificially created for a certain purpose languages are called “constructed languages“. There are over 1000 such languages and growing. Some examples: Esperanto, Lojban, Toki Pona, Elf, etc.

Natural Language Processing (NLP) is a subfield of Artificial Intelligence and Mathematical Linguistics. It studies computer analysis and synthesis of natural languages. Applied to artificial intelligence, analysis means understanding of language and synthesis means generation of grammatically correct text. Solving these problems leads to creation of more efficient means of interaction between human and machine.
An example of a dialog:
(1) What is a natural language?
(1) Natural language (NL) is a language that people use to talk which was not created artificially.
(Natural language (NL) is a language that people use to talk which was not created artificially.)
(2) Name some examples of natural languages.
(2) Russian, English.
(Examples of natural languages are Russian, English, Chinese, Kazakh, etc.)
(3) Can you name more?
(3) Chinese, Kazakh.
(Examples of natural languages are Russian, English, Chinese, Kazakh, etc.)
(4) Can you name some formal languages?
(4) Mathematical logic; programming languages; languages, generated by regular expressions, finite-state machines, Chomsky hierarchy.
(Examples of formal languages: mathematical logic; programming languages; languages, generated by regular expressions, finite-state machines, Chomsky hierarchy, etc.)
(5) What are artificially created languages called?
(5) Constructed languages.
(Artificially created for a certain purpose languages are called ”constructed languages”.)
(6) Can you name some constructed languages?
(6) Esperanto, Lojban, Toki Pona, Elf.
(Some examples: Esperanto, Lojban, Toki Pona, Elf, etc.)
(7) Can you name the subfield of Artificial Intelligence and Mathematical Linguistics?
(7) Natural Language Processing.
(Natural Language Processing (NLP) is a subfield of Artificial Intelligence and Mathematical Linguistics.)
(8) What do ”analysis” and ”synthesis” mean when applied to artificial intelligence.
(8) Analysis means understanding of language and synthesis means generation of grammatically correct text.
(Applied to artificial intelligence, analysis means understanding of language and synthesis means generation of grammatically correct text.)
(9) Is it possible to create more efficient means of interaction between human and machine.
(9) Yes, by solving the analysis and synthesis problems.
(It studies computer analysis and synthesis of natural languages. Applied to artificial intelligence, analysis means understanding of language and synthesis means generation of grammatically correct text. Solving these problems leads to creation of more efficient means of interaction between human and machine.)
(10) How many constructed languages there are?
(10) Over 1000.
(Artificially created for a certain purpose languages are called ”constructed languages”. There are over 1000 such languages and growing.)
(11) Are there new constructed languages being developed?
(11) Yes.
(Artificially created for a certain purpose languages are called ”constructed languages”. There are over 1000 such languages and growing.)

B. Methods

We consider a hybrid solution, combining both rule-based and machine learning-based approaches to question generation. The rule-based system uses a set of handcrafted linguistic rules described by a formal grammar with additional syntactic conditions. However, the grammar is not evaluated strictly: the analyzer can omit parts of a sentence if it is determined inessential to the overall meaning of the sentence. We can summarize the algorithm the following way:
1. The analyzer determines the set of rules suitable for processing the current sentence. The set of rules is determined based on keywords (sequences of terminal symbols) contained in each rule.
2. A rule, which consists of terminal and nonterminal symbols, is projected onto the current sentence; the positions of terminals are used to calculate temporary values for all nonterminals in the rule.
3. The algorithm is repeated for the temporary value of every nonterminal if the grammar has a rule for the nonterminal.
4. If there are no rules for the nonterminal, the temporary value



Did you know that Albert Einstein could not speak until he was four years old, and did not read until he was seven? His parents and teachers worried about his rnenta1ability.

Beethoven's music teacher said about him,"As a composer he is hopeless." What if this young boy believed it?

When Thomas Edison was a young boy,his teachers said he was so stupid that he could never learn anything.He once said,"I remember I used to never be able to get along at schoo1.I was always at the foot of my class...My father thought I was stupid,and I almost decided that l was a stupid person."What if young Thomas believed what may said about him?

When the sculptor Auguste Rodin was young; he had difficulty learning to read and write.:. Today, we may say he had a learning disability. His father said of him, "I have an idiot for a son. "His uncle agreed. "He's uneducable," he said. What if Rodin had doubted his ability?

Walt Disney was once fired by a newspaper editor because he was thought to have no "good ideas". Enrico Caruso was told by one music teacher, "you can't sing. You have no voice at all. " And an editor told Louisa May Alcott that she was unable to write anything that would have popular appeal.

What if these people had listened and become discouraged? Where would our world be without the music of Beethoven, the art of Rodin or the ideas of Albert Einstein and Walt Disney? As Oscar Levant once said, "It's not what you are but what you don't become that hurts. "

You have great potential.When you believe in all you can be, rather than all you cannot become, you will find your place on earth.

Знаете ли вы, что Альберт Эйнштейн не мог говорить до четырех лет и не читал до семи? Его родители и учителя беспокоились о его умственных способностях.

Учитель музыки Бетховена говорил о нем: "как композитор Он безнадежен.- А что, если этот мальчик поверит в это?

Когда Томас Эдисон был маленьким мальчиком,его учителя говорили, что он настолько глуп, что никогда ничему не научится.Однажды он сказал: "Я помню, что раньше никогда не мог ладить в шклолу. Я всегда был самым слабым учеником в классе...Мой отец считал меня глупым, и я почти решил, что я был глупой личностью.-А что, если юный Томас поверил бы тому, что о нем говорили?

Когда скульптор Огюст Роден был молод, он с трудом учился читать и писать.:. Сегодня мы можем сказать, что у него была неспособность к обучению. Его отец сказал о нем: "у меня сын-идиот. - Его дядя$$$$согласился. -Он необразован, - сказал он. Что, если Роден усомнился в своих способностях?

Уолт Дисней однажды был уволен редактором газеты, потому что считалось, что у него нет "хороших идей". Один учитель музыки сказал Энрико Карузо: "ты не умеешь петь. У тебя вообще нет голоса. - И редактор сказал Луизе Мэй Элкотт, что она не может написать ничего, что могло бы привлечь внимание публики.

Что, если бы эти люди послушались и впали в уныние? Где был бы наш мир без музыки Бетховена, искусства Родена или идей Альберта Эйнштейна и Уолта Диснея? Как однажды сказал Оскар Левант: "больно не тому, кто ты есть, а тому, кем ты не становишься. "

У вас большой потенциал.Когда вы верите во все, чем вы можете быть, а не во все, чем вы не можете стать, вы найдете свое место на земле.

(1) - What age did Einstein start talking?
(1) - Four
(know that Albert Einstein could not speak until he was four years old,)
(2) - When did he start reading?
(2) - Seven
(did not read until he was seven? His parents and teachers worried about his rnenta1ability.)
(3) - Who was worried about him?
(3) - His parents
(His parents and teachers worried about his rnenta1ability.)
(4) - Did Beethoven's music teacher support him?
(4) - No
(Beethoven's music teacher said about him,"As a composer he is hopeless." What if this young boy believed it?)
(5) - Did Thomas Edison's teachers think he was smart?
(5) - No
(When Thomas Edison was a young boy,his teachers said he was so stupid that he could never learn anything.)
(6) - Did Thomas have any family members that thought he was stupid?
(6) - Yes
(When Thomas Edison was a young boy,his teachers said he was so stupid that he could never learn anything.He once said,"I remember I used to never be able to get along at schoo1.I was always at the foot of my class...My father thought I was stupid,and I almost decided that l was a stupid person."What if young Thomas believed what may said about him?)
(7) - Which one in particular?
(7) - His father
(My father thought I was stupid,and I almost decided that l was a stupid person."What if young Thomas believed what may said about him?)
(8) - What did Auguste Rodin become famous for?

(1) - в каком возрасте Эйнштейн начал говорить?
(1) - в четыре года
(знайте, что Альберт Эйнштейн не мог говорить, пока ему не исполнилось четыре года,)
(2) - когда он начал читать?
(2) - в семь лет
(не читал до семи лет? Его родители и учителя беспокоились о его умственных способностях.)
(3) - кто беспокоился о нем?
(3) - его родители
(Его родители и учителя беспокоились о его умственных способностях)
(4) - поддерживал ли Бетховена его учитель музыки ?
(4) - Нет
(Учитель музыки Бетховена сказал о нем: "как композитор Он безнадежен.- А что, если этот мальчик поверит в это?)
(5) - учителя Томаса Эдисона считали его умным?
(5) - Нет
(Когда Томас Эдисон был маленьким мальчиком,учителя говорили, что он настолько глуп, что никогда ничему не научится.)
(6) - были ли у Томаса родственники, которые считали его глупым?
(6) - Да
(Когда Томас Эдисон был маленьким мальчиком,учителя говорили, что он настолько глуп, что никогда ничему не научится.Однажды он сказал: "Я помню, что раньше никогда не мог ладить в шклолу. Я всегда был самым слабым учеником в классе... Мой отец считал меня глупым, и я почти решил, что я был глупой личностью.-А что, если юный Томас поверил бы тому, что о нем говорили?)
(7) - Кто именно?
(7) - его отец
(Мой отец считал меня глупым, и я почти решил, что я был глупой личностью.-А что, если юный Томас поверил бы тому, что о нем говорили?)

Обработано: 11

Статус фиксации: сохранено

Осталось 290

☑ Отредактированный перевод

Fig. 1. Example of a conversation

is stored and then retrieved if the nonterminal corresponds to a semantic relation.

5. If some rules exist for the nonterminal and they define the possible parts of speech the nonterminal may contain, the temporary value is compared to all permissible combinations of parts of speech. The nonterminal is retrieved if at least one combination is found; otherwise, it is inferred that the temporary value can not be used to describe the nonterminal and the algorithm restarts.

Since the data is heterogenous, we find it practical to use a simple method — developing a more complicated grammar would require more time and effort while not necessarily guarantee better performance.

This work employs information extraction techniques based on lexico-syntactic templates and rules describing term structure, as well as common usage patterns in Russian scientific texts. Our approach necessitates involvement of an expert creating templates/rules for information extraction. We attempted to extract terms, synonyms, hyponyms, instances, and attributes.

The patterns used for extracting different objects were segregated into several groups which are reviewed below. Naturally, some of the candidate terms, synonyms and hyponyms are incorrect; we can tailor a list of stopwords to weed out some of the false positives.

A necessary step for applying information extraction algorithms is text preprocessing [39], [40]. Aside from removing stopwords, we removed punctuation and other irrelevant symbols; then we performed tokenization. As some extracted terms may be inflected, we also used lemmatization to normalize extracted words as postprocessing. Here are some common patterns we noted to consider for further work on extracting terms:

1) a term may be at the start of a new line;
2) a term may be written in uppercase;
3) a term may be divided by symbols that do not pertain to the term itself;
4) a word or a phrase that is part of a term may not be in its base form;
5) a term and its definition may be separated by the term's translation in parentheses; we consider that a synonym, too.

The principle for extracting hyponyms and synonyms is analogous to that for extracting terms except for the created rules. The templates we used to extract different terms are divided into groups as follows (Tables II and III):

Terms were extracted for constructing questions using different types of templates. Here are some examples of extracted questions:

As a machine learning-based method we use the Reinforced Dynamic Reasoning network suggested by Pan et al. which represents an encoder-decoder approach. The encoder, a bidirectional LSTM, iteratively reads the conversation history C, encoding it jointly with the rationale R through the use of an alignment matrix: $S = R^T C$. The representations of the conversation history and the rationale are obtained via $H = R{\cdot}\mathrm{softmax}(S)$ and $A = C{\cdot}\mathrm{softmax}(\mathrm{S})$, respectively. The final representation is given by $G = [C; H] \cdot \mathrm{softmax}(S^T)$ where $[;]$ means row-wise concatenation. G, in turn, is fed into the

TABLE II. RULE GROUPS

Rule for extraction, **Rule examples**	**Example in a sentence**
Of terms TERM is . . . ; TERM is called . . . ; TERM is performed by ...; . . . is called TERM	”Graphemics analysis is the first step in automated natural language processing”. In this sentence, the term is ”graphemics analysis”.
Of hyponyms TERM consists of HYPONYM1, HYPONYM2, . . . ;TERM are HYPONYM1, HYPONYM2, . . . ;TERM: HYPONYM1, HYPONYM2, . . .	”Examples of formal languages: mathematical logic; programming languages; languages, generated by regular expressions, finite-state machines, Chomsky hierarchy, etc”. Here, ”formal languages” is a term and ”mathematical logic”, ”programming languages”, ”languages, generated by regular expressions”, ”finite-state machines”, and ”Chomsky hierarchy” are hypomyns.
Of synonyms TERM or SYNONYM . . . ; TERM (SYNONYM) . . . ; TERM . . . also called SYNONYM . . .	”Natural Language Processing (NLP) is a subfield of Artificial Intelligence and Mathematical Linguistics”. ”Natural Language Processing” and ”NLP” are synonyms.

TABLE III. TEMPLATES FOR QUESTIONS AND ANSWERS

Template for question and answer	**Question and answer in natural language**
What is TERM? It is TERM	What is a natural language? A language used for communication between people.
What kinds of TERM there are? HYPONYM(TERM)	What kinds of languages there are? Natural language, formal language, constructed language.
What are some examples of TERM? EXAMPLE(TERM)	What are some examples of the Elf language? Sindarin, Quenya.
What TERM consists of? HYPONYM (TERM)	What a language analysis consist of? Graphemics analysis, morphological analysis, fragmentation analysis.

integration network, another bi-directional LSTM, returning a list of vectors $U^0 = [u_1^0, u_2^0, ...]$. The decoder, an LSTM network with an MLP layer as head, reads that list iteratively, then generates the question by sampling word probabilities.

The rationale is extracted from the source text. Pan et al. propose picking whole sentences consecutively. We also try extracting sentence clauses to use as rationales. Most often rationale is a standalone entity (a name, date, place, etc.) or with some context, sometimes forming a clause. Rationale can also be a sentence or multiple sentences. We noted that the authors' strategy for extracting rationale is not perfectly followed. The model takes a conversational context as input (questions with answers) and a rationale which should contain information to facilitate a new question, as well as enough context for the copying mechanism to work correctly. We tried three strategies for updating rationales:
1) leave the original rationale;
2) use the whole sentence the original rationale came from;
3) use the clause containing the original rationale or the entity from the answer. The first strategy allows us to follow more closely along with the authors' experiment; however, the lack of context hinders the quality of machine translation. This approach also limits the way we are allowed to prepare rationales when using the system to generate questions. The second strategy is easier to implement but gives way to confusion as to which entity the question should be generated for. Finally, the third strategy seems to lack most of the disadvantages; still, clauses are not always straightforward to extract.

For training the network, we use two datasets: an English CoQA and a Russian translation of **CoQA**.

Statistics for both datasets can be seen in Table IV.

TABLE IV. DATA STATISTICS

	CoQA-RU	**CoQA-EN**
Dialogue count	5656	7146
Min dialogue length	1	1
Average dialogue length	17.8	36
Max dialogue length	36	14.8
QA pair count	83584	105766
Min question length (in tokens)	1	1
Average question length	5.7	6.5
Max question length	62	48
Min rationale length	1	1
Average rationale length	11.2	10.3
Max rationale length	385	422
Min answer length	1	1
Average answer length	2.8	2.9
Max answer length	385	422

For tokenization, we used traditional word-by-word tokenization, as well as the Byte-Pair Encoding Tokenizer (https://opennmt.net/OpenNMT/tools/tokenization/#bpe), which allows learning embeddings for sub-word tokens, providing the benefit of retrieving embeddings for new words. We try several approaches to text generation: beam search with copying, top-k sampling, and temperature sampling.

IV. Results

Comparative analysis of the authors' system of rules for extracting terms (Rule-Based Term Extraction) and the most common libraries for extracting terms from texts in Russian are presented in Table V. The Summa library is based on a simplified graph model that allows to summarize text, extract keywords, and terms. For experimenting with a rule-based approach we used the **rutermextract** library; for an ML-based alternative, we used **gensim**.

TABLE V. Performance for term extraction

Algorithm	Precision	Recal	F1-score
Rutermextract	0,134	0,9	0,23
Gensim	0,269	0,566	0,35
Summa	0,306	0,633	0,41
Rule-Based Term Extraction	0,92	0,8	0,85

The set of rules we developed for extracting terms surpasses the results we got using other tools significantly. Table VI presents the results for extracting synonyms, definitions, and hyponyms from textual materials; since there are no solutions available for extracting these types of relations in Russian, no comparison may be drawn.

TABLE VI. Performance for extraction of other types of relations

Objects	Precision	Recall	F1-score
Terms	0,92	0,8	0,85
Definitions	0,66	0,53	0,58
Hyponyms	0,88	0,47	0,55
Synonyms	0,916	0,785	0,84

The formal knowledge base formed using the extracted terms, definitions, synonyms and hyponyms allowed us to create a sufficiently large set of 'oral exam' questions and correct answers, such as:

[('what kinds of languages there are?', 'natural language, formal language, constructed language')]

('what is a formal language?', 'a language created with a certain goal')

('what a language analysis consist of?', 'graphemics analysis, morphological analysis, fragmentation analysis')

('what language synthesis is part of?', 'natural language processing')

('what are some morphological characteristics?', 'gender, grammatical number, case, declension, and tense')

A. Experiments with ReDR

The best performance was shown by the model simultaneously trained on both languages. In this case, overfitting happens way later and the generated questions make sense. Most frequent questions are about terms and definitions.

Only the training in Russian suffers from early overfitting.

Syntagmas extraction for rationales drops the BLEU score significantly (Fig. 2).

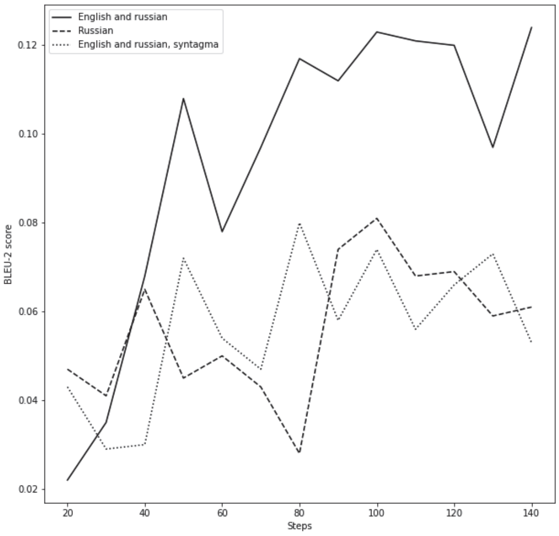


Fig. 2. BLEU score

Forced generation of longer questions decreases the number of short questions (Where? How much? What is next?) but the generated questions most often have unacceptable quality. It seems that the model can not connect a context with a rationale and chooses to generate a general question.

Balancing the dataset for most general and short questions decreases the number of them at generation but does not improve the BLEU score since more "broken" questions appear.

BLEU evaluation is not perfect for this task. For the same context and rationale, it is possible to generate multiple questions. Often, while a number of correct questions are generated, only a single one is considered a hit. However, this is mostly an issue with the dataset (a dataset with multivariate questions and answers would be significantly more helpful).

Further improvements to the model can be achieved in several ways:

1) By including an additional label that marks the type of a question, its template, and the interrogative word (most of the questions are generated with "what").

2) By implementing a rule-based algorithm that can separate rationales and questions by terms (objects) and for each question-answer pair determine the target object. It makes the approach more agile and there is a hypothesis that it can improve the performance of the method overall.

3) Since the datasets are rather small for the model to learn the language, transfer learning can be applied to generate a pre-trained transformer.

After assessing the generated questions manually, we came to the following conclusions:

1) The quality of the generated questions does not seem to be as high as for those generated by Pan et al. [36].

2) The model overfits quickly, however, the early stages of overfitting are when the questions become more diverse, such as:

- What is a language? ("language" being the correct entity from the rationale)
- What is that?
- Anything else?
- How many was that?
- Why?

Early overfitting also occurs when the model starts to pay attention to the context:

- How many levels? [of sentence analysis]
- What happened to [term] [gibberish]?
- What is information?
- What is about [incongruity] lexical resource?

3) Reinforcement learning performed in the same manner as Pan et al. did not seem to improve questions generated by our model.

V. Conclusion

The ReDR model could be improved by allowing an expert to point out the entity the generated question should be directed at. The model also seems to get confused when the same rationale is being used several times in a row, even if each question points at a different aspect — this could potentially be alleviated by incorporating information about the type of question to be generated. Finally, early overfitting leads to the conclusion that available data is still insufficient; additional data can be gathered after the system's deployment. Using a larger dataset will likely boost the quality of generated questions. In General, the hybrid approach presented in this paper allows to vary the number of questions generated by the neural network and when obtaining higher results, the proportion of such questions will increase.

Acknowledgment

This work was partially financially supported by the Government of the Russian Federation (Grant 08-08).

References

[1] N. Shilkina, A. Maltseva, O. Makhnytkina, M. Titova, E. Gubernatorova, I. Katsko, F. Mirzabalaeva, and S. Shusharina, "Social media as a display of students' communication culture: case of educational, professional and labor verbal markers analysis," *Communications in computer and information science*, pp. 384–397, 2019.

[2] M. Heilman, "Automatic factual question generation from text," 2011.

[3] M. H. Noah A. Smith, "Question generation via overgenerating transformations and ranking," pp. 1–15, 2009.

[4] E. Bol'shakova and K. Ivanov, "Selecting terms and their relationships for the subject index of a scientific text," *"Sixteenth national conference on artificial intelligence"*, pp. 253–261, 2018.

[5] M. M. Yakupova, "Automatic extraction of terms from messages," *Science, technology and education*, pp. 61–64.

[6] M. Zagorulko and E. Sidorova, "The extraction system of the subject terminology from the text based on lexico-syntactic patterns," *Proceedings of the XIII International conference " Problems of control and modeling in complex systems*, pp. 506–511, 2011.

[7] J. Foo, "Term extraction using machine learning," pp. 1–8, 12 2009. [Online]. Available: https://cl.lingfil.uu.se/~nivre/gslt/foo.pdf

[8] M. Arun S., V. Dale, and W. Andrew, "Mining measured information from text," *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 5–10, 6 2015.

[9] L. Nikola, F. Darja, and E. Tomaž, "Kas-term: Extracting slovene terms from doctoral theses via supervised machine learning," *Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science, vol 11697*, pp. 115–126, 2019.

[10] H. H. Peng Li, "Clinical information extraction via convolutional neural network," *CoRR*, vol. abs/1603.09381, 3 2016. [Online]. Available: http://arxiv.org/abs/1603.09381

[11] Y. Luan, M. Ostendorf, and H. Hajishirzi, "Scientific information extraction with semi-supervised neural tagging," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2641–2651, 9 2017.

[12] N. Nityasya, R. Mahendra, and M. Adriani, "Hypernym-hyponym relation extraction from indonesian wikipedia text," *2018 International Conference on Asian Language Processing (IALP)*, pp. 285–289, 11 2018.

[13] K. Sabirova and A. Lukanin, "Automatic extraction of hypernyms and hyponyms from russian texts," *CEUR WORKSHOP PROCEEDINGS 3. 2014*, pp. 35–40, 2014.

[14] G. Kurdi, J. Leo, B. Parsia, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, 11 2019.

[15] K. Kettip and A. Wangperawong, "Question generation by transformers," *arXiv preprint arXiv:1909.05017*, 2019.

[16] S. Santhanam, "Context based text-generation using lstm networks," *arXiv:2005.00048*, 2020.

[17] K. Xiao, X. Zhou, Z. Wang, X. Duan, and Zhang, "Question generation based product information," *Natural Language Processing and Chinese Computing. NLPCC 2019. Lecture Notes in Computer Science, vol 11839*, pp. 445–455, 2019.

[18] J. . L. M. Chen, Meixi Zhao, "Using multiple encoders for chinese neural question generation from the knowledge base," *IOP Conference Series: Materials Science and Engineering.*, pp. 1–6, 2019.

[19] M. Heilman and N. A. Smith, "Question generation via overgenerating transformations and ranking," *IOP Conference Series: Materials Science and Engineering.*, 2009.

[20] N. Duan†, D. Tang†, P. Chen, and M. Zhou†, "Question generation for question answering," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 866–874, 2017.

[21] K. Payal, R. Konigari, H. Mukul, and S. Manish, "Automatic question generation using relative pronouns and adverbs," *Proceedings of ACL 2018, Student Research Workshop*, p. 153–158, July 2018.

[22] S. Tarasenko and N. Ryazanova, "Analysis of methods for automatic generation of questions in natural language," *Engineering journal*, pp. 1032–1037, 2015.

[23] W. Zhou, M. Zhang, and Y. Wu, "Question-type driven question generation," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[24] A. Killawala, I. Khokhlov, and L. Reznik, "Computational intelligence framework for automatic quiz question generation," 07 2018, pp. 1–8.

[25] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," 01 2017, pp. 1342–1352.

[26] Y. Zhao, X. Ni, Y. Ding, and Q. Ke, "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*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3901–3910. [Online]. Available: https://www.aclweb.org/anthology/D18-1424

[27] T. Scialom, B. Piwowarski, and J. Staiano, "Self-attention architectures for answer-agnostic neural question generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019,

pp. 6027–6032. [Online]. Available: https://www.aclweb.org/anthology/P19-1604

[28] N. Duan, D. Tang, P. Chen, and M. Zhou, “Question generation for question answering,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 866–874. [Online]. Available: https://www.aclweb.org/anthology/D17-1090

[29] J. Cho, M. Seo, and H. Hajishirzi, “Mixture content selection for diverse sequence generation,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3121–3131. [Online]. Available: https://www.aclweb.org/anthology/D19-1308

[30] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon, “Unified language model pre-training for natural language understanding and generation,” *CoRR*, vol. abs/1905.03197, 2019. [Online]. Available: http://arxiv.org/abs/1905.03197

[31] Y. Yan, W. Qi, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, “Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training,” 2020.

[32] S. Reddy, D. Chen, and C. D. Manning, “Coqa: A conversational question answering challenge,” *Transactions of the Association for Computational Linguistics*, vol. 7, p. 249–266, Mar 2019. [Online]. Available: http://dx.doi.org/10.1162/tacl_a_00266

[33] Y. Gao, P. Li, I. King, and M. R. Lyu, “Interconnected question generation with coreference alignment and conversation flow modeling,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4853–4862. [Online]. Available: https://www.aclweb.org/anthology/P19-1480

[34] B. Pan, H. Li, Z. Yao, D. Cai, and H. Sun, “Reinforced dynamic reasoning for conversational question generation,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/p19-1203

[35] A. Matveev, O. Makhnytkina, I. Lizunova, T. Vinogradova, A. Chirkovskii, A. Svischev, and N. Mamaev, “A virtual dialogue assistant for conducting remote exams.” *Proceedings of the 26th Conference of Open Innovations Association FRUCT*, pp. 284–290, 2020.

[36] B. Pan, H. Li, Z. Yao, D. Cai, and H. Sun, “Reinforced dynamic reasoning for conversational question generation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2114–2124. [Online]. Available: https://www.aclweb.org/anthology/P19-1203

[37] V. Krasavina and A. Mirzagitova, “Search optimization in the system leadscanner with automatic extraction of key words and phrases,” *Proceedings of the international conference ”Corpus linguistics-2015*, p. 296306, 2015.

[38] A. Moskvina, O. Mitrofanova, A. Erofeeva, and Y. Harabet, “Automatic selection of keywords and phrases from russian-language text corpora using the rake algorithm,” *Proceedings of the international conference ”Corpus linguistics-2017*, p. 268277, 2015.

[39] E. Klyshinskij and N. Kochetkova, “Method for extracting technical terms using the oddity measure,” *New information technologies in automated systems*, pp. 365–370, 2014.

[40] S. Pivovarova, “Identifying candidate terms for a multilingual terminology dictionary,” *Collection of scientific articles of the XIX joint conference” Internet and modern society ” IMS-2016*, pp. 55–64, 2016.