Using Hybrid Intelligent Information System Approach for Text Question Generation

Marina Belyanova[a], Valeriy Chernenkiy[a], Yuriy Kaganov[a] and Yuriy Gapanyuk[a]

[a] *Bauman Moscow State Technical University, 2-ya Baumanskaya ul., 5, Moscow, 105005, Russian Federation*

Abstract

The paper is dedicated to the question generation based on the text information. The review of modern approaches to question generation from text is provided in the paper. It is noted that the question generation approaches are divided into two categories: the logical approach and the machine learning approach. The architecture of the information system for question generation is proposed, based on the HIIS conception. The metagraph embedding and storage structure details are discussed. The experimental results show the applicability of the proposed approach.

Keywords 1

Hybrid intelligent information system (HIIS), the consciousness of the information system, the subconsciousness of the information system, overgenerate-and-rank, sequence-to-sequence, semantic hypergraphs, natural language processing, question generation.

1. Introduction

The task of automatic question generation from the text may be used for the automatization of quality assurance for the understanding of the read text by the students. The module of this kind can also improve chat-bot systems in the beginning, continuing or enriching the dialogue.

Asking questions is an essential ability of the human brain that affects the learning process and improves it. In [1], it is stated that in order to use information in more complex situations, it is essential to extract it from the text.

In order to ask the question right, it is crucial to be familiar with the context of the question: the information field in which the question appears. Thus, in the case of the task of question generation from the text, the context may be presented by the text itself or by various ontological characteristics of the words and phrases in it.

The motivation for the idea of the testing materials generation task is to reduce the amount of manual work and time that is spent on making the tests for the text understanding check. On the other hand, there is also the motivation to enrich the chatting techniques in the dialog systems to provide better user experience for future communication with artificial intelligence.

It is stated that the beginning of the researches in the automatic question generation is 1970-s, and by now, this area is rapidly developing due to the process of adding online-technologies to the education process.

The relevance of the generating quality check materials is that it would decrease the time for making a large number of test tasks for checking of how students learn the material. The relevance for improving the techniques for dialogue systems is driven by the need to enrich the communication for artificial intelligence.

Russian Advances in Fuzzy Systems and Soft Computing: selected contributions to the 8-th International Conference on Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT-2020), June 29 – July 1, 2020, Smolensk, Russia
EMAIL: kaganov.y.t@bmstu.ru(Yu.T.Kaganov)

2. The logic-based and machine learning-based approaches to the question generation from the text

There can be two different categories on which the approaches for question generation from the text are divided.

In the first type, the question is generated from the text, based on the manually provided rules that map given sentences to the question text. That may require knowledge of the text language structure, basically the knowledge of how the question is formed from the declarative sentence.

For the second category, the learning on the huge amount of texts with the pre-written questions is required, which provides a flexible generalization of question generation rules, but may take a large amount of time for learning and collecting the training data.

2.1. Logic-based approach

The logic-based approach separates into two groups: first is so-called content selection, when the model selects the part of the text and, based on this part of the text, determines what particular kind of the question is required for this part. The second is question construction. It converts the intermediate representation (such as an ontological structure) to a natural language question.

As an example of first category task solving, in the research [2], the mapping from the text to the question is made using the particular set of the rules: words from the sentence were tagged with parts-of-speech tag and, depending on which predefined part of speech structure was found in the sentence, it was rebuilt to get the questions. The authors used an array of predefined regexps to reconstruct the text question.

In work [3], the question generation is made on the ontological structure, such as the phylogenetic tree of species, presented as a graph, where the vertices are the titles of the species' category, and the connections are the relations between categories. Using the "expanding" of the categories titles and their relationships with the hand-written rules for question generation, the authors generated the questions that were used for the checking of structural knowledge of the topic. It can be an example of the second (question construction) approach.

In order to reduce the necessity to build the inference rules by themselves manually, in [4] the strategy called "overgenerate-and-rank" was used. First, the text was summarized into several simple facts. Then these facts were transformed to the question and ranked by the model, trained on manually written text.

2.2. Machine learning-based approach

For machine learning-based approaches, nowadays, the Seq2Seq (sequence-to-sequence) approach, which allows the sequence generation based on the given sequence, is gaining popularity. It is inspired by the task of machine translation.

In this approach, given a passage X = (x_1, x_2, …, x_n) and a target answer A as an input, the model aims to create a question Y=(y_1, y_2, …, y_m) asking about the target answer A, so that the search of the best question $\bar{Y}$ maximized the conditional likelihood given the passage X and the answer A:

$$\bar{Y} = arg \max_{Y} P(Y|X, A) = arg \max_{Y} \sum_{t=1}^{m} P(y_t|X, A, y < t).$$

There are different datasets that are used to evaluate machine-learning models, such as SQuAD, NewsQA, MS MARCO, NarrativeQA. The difference between them can be in the type of the answer (it can be human-generated or taken from the given text), in the domain area (Wikipedia, web articles of the story), and the amount of data that is stored in it.

The approach to generate questions based on the text and an answer with an additional reinforcement learning is described in the paper [5]. In this research, the sentence and the text of the answer are encoded into the graph form with the vertices, representing the word role and position in the text, and the edges, representing the relations between words and sentences. The output of the model is the text of the question. The input of the sentence in this paper is encoded with the bidirectional graph neural network (BiGNN) [6]. This hack allowed authors to encounter not only forward but also backward dependencies in the sentence.

SQuAD Benchmark dataset (split-1 and split-2) was used as the training and testing set. This dataset was transformed into the graph form and is stored in it. It is based on the manually tagged questions and answers of the Wikipedia articles.

The approach for text tagging and providing text to the machine learning algorithm in the graph form called "semantic hypergraphs" is described in the paper [7]. By this approach, the authors wanted to solve several problems: first, to provide the visualization of the text that allows saving recursive connections between sentence parts. Second, to provide the mapping that allows storing connections that would disappear if the syntactic tree structure is used for the text presentation instead of the more general graph form. New attributes could also increase the quality of models that are trained using them, as stated by the authors, but they do not provide the results of model training using their data structure.

In this paper, as the graph structure of the text given to the neural network model, the metagraph structure is proposed. Metagraphs allow building multilevel ontological constructions to extract additional data about the entities from the text and enrich the sentence structure with extra features.

3. The HIIS-based architecture of the proposed system

The architecture of the proposed system is based on the HIIS (Hybrid Intelligent Information System) approach [9]. The generalized HIIS architecture represented in Figure 1 includes the following components: the environment (ENV); the subconsciousness module (MS); the consciousness module (MC); the boundary model of consciousness and subconsciousness (BOUND), the module of interaction (MI).

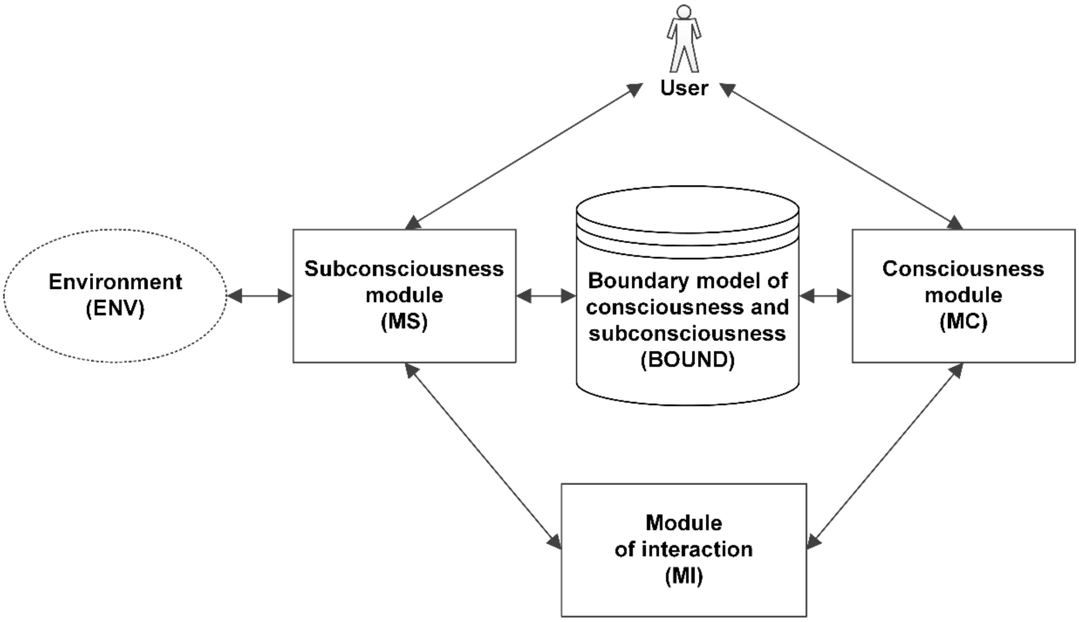
User
Environment (ENV)
Subconsciousness module (MS)
Boundary model of consciousness and subconsciousness (BOUND)
Consciousness module (MC)
Module of interaction (MI)

Figure 1: The generalized HIIS architecture

The subsystem of subconsciousness is related to the environment in which a HIIS operates. Because the environment can be represented as an unstructured data or a set of continuous signals, the data processing techniques of the MS are mostly based on neural networks, fuzzy logic, and combined neuro-fuzzy methods.

The subsystem of consciousness is based on conventional data and knowledge processing, which may be based on traditional programming or workflow technology. As to the data model, the MC uses ontology-based models. They can be classical ontologies, which are developed within the Semantic

Web technology (RDF, RDFa, OWL, and OWL2 standards), or nonstandard ontology models. Also, the classical object-oriented approach, which in practice is used in most information systems, may be included.

The boundary model of consciousness and subconsciousness is intended for deep integration of modules of consciousness and subconsciousness and represents an interface between these modules with the function of data storage.

The main task of the subconsciousness module is to recognize elements of ontology from the environment. If we consider the consciousness module as a kind of expert system, then the recognized elements of the ontology can be considered as elements of the operating memory of the expert system that trigger the corresponding rules. Depending on the goals of the system, rules can generate output information for the user or signals for the subconscious module that have the desired effect on the environment.

From the interaction point of view, the following options or their combinations are possible in a HIIS:

- Interaction is implemented in the environment. The MS reads the data from the environment, converts them, and transmits them to the MC. The MC performs logic processing and returns the results to the MS. The MS writes the results into the environment, where they can be read by another HIIS.
- The MI is used for the interaction with another HIIS. Depending on the tasks to be solved, the MI can interact with the MC (which is typical for conventional information systems) or with the MS (which is typical for systems based on soft computing).
- User interaction can be carried out using the MC (which is typical for conventional information systems) or through the MS (which can be used, for example, in automated simulators).

The proposed HIIS architecture is considered as a generalized approach that should be adapted to create information systems in specific subject areas. The architecture of the proposed system based on the HIIS concept is presented in Figure 2.

The environment in the proposed system is text documents used to generate questions.

The subconsciousness module of the system (MS) includes a “module for the formation of concepts” MS.1, a “module for generating layouts of questions” MS.2, and the “embedding module” MS.3.

The “module for the formation of concepts” implements the selection of concepts and the relationships between them from the source text.

The “module for generating layouts of questions” uses machine learning techniques for question generation. This module can simultaneously call several methods for generating question layouts and create several question variants. Therefore, the double arrow in the output is used.

The “embedding module” is used for creating vector representations of the text and graph (metagraph) fragments. The vector representations are based on the MS.1 and MS.2 modules outputs.

The boundary model of consciousness and subconsciousness contains concepts selected from the text and links between them, generated question layouts, vectorization results. The storage is implemented on the basis of a metagraph data model, which allows semantic enrichment of concepts, as well as linking concepts with fragments of generated question layouts.

The consciousness module of the system (MC) includes the described below modules.

The “module for semantic enrichment” MC.1 is used to expand concepts useful for the formation of questions with information based on dictionaries, thesauri, and text corpora.

The “correction of layouts of questions” module MC.2 is used to correct the generated question layouts based on dictionaries, thesauri, text corpora (in text and embedded forms): combating typos, improving the wording, eliminating word inconsistencies in the text.

The “logical formation of questions module” MC.3 is used to form questions based on enriched concepts.

The “machine learning formation of questions module” MC.4 is used to form questions based on machine learning, deep learning, reinforcement learning, etc.

The “hybrid formation of questions module” MC.5 is used to generate questions both on the basis of enriched concepts and on the basis of machine learning methods.

The “quality assessment module” MC.6 compares the quality of generated questions based on quality metrics. Variants of the generated questions with quality metrics are issued to the user.

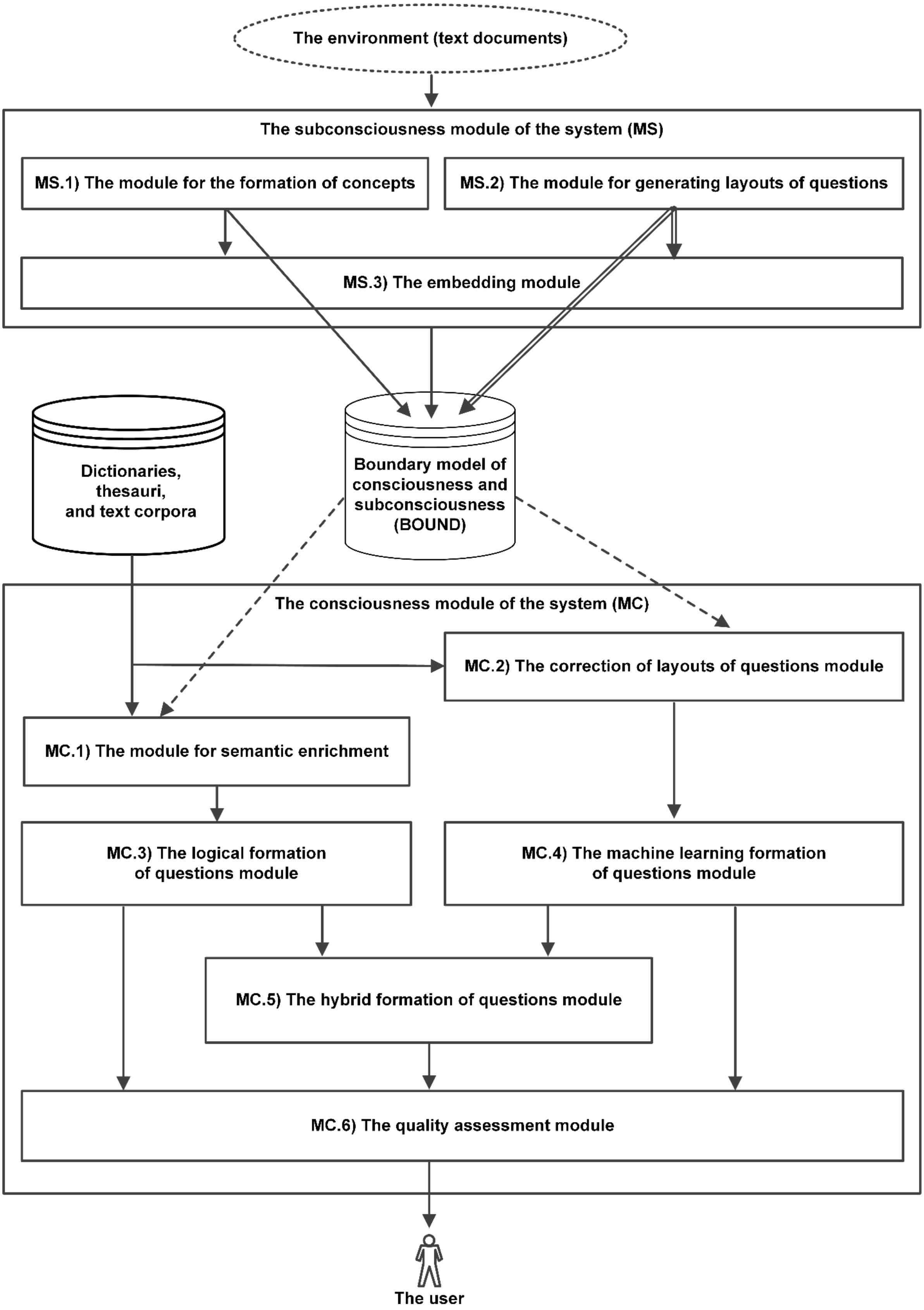
The environment (text documents)
The subconsciousness module of the system (MS)
MS.1) The module for the formation of concepts
MS.2) The module for generating layouts of questions
MS.3) The embedding module
Dictionaries, thesauri, and text corpora
Boundary model of consciousness and subconsciousness (BOUND)
The consciousness module of the system (MC)
MC.2) The correction of layouts of questions module
MC.1) The module for semantic enrichment
MC.3) The logical formation of questions module
MC.4) The machine learning formation of questions module
MC.5) The hybrid formation of questions module
MC.6) The quality assessment module
The user

Figure 2: The architecture of the proposed system

The use of all the modules may be redundant. Let’s consider special cases of the architecture of the system, which use only some modules.

- The classical architecture of concept-based question generation includes a chain of modules: MS.1-MC.3. In this case, concepts are extracted from the text, and questions are generated based on rules. The main disadvantage of this option is the great complexity of developing a system of rules.
- The classical architecture of generating questions based on machine learning methods includes a chain of modules: MS.2-MC.4. In this case, machine learning methods are used to generate questions, and then a quality assessment is performed. The main disadvantages of this option are inherent in almost all machine learning methods: lack of explanation (the model cannot explain why it generated such a question), the inability to correct the algorithm for generating questions using rules.
- MS.1-MC.1-MC.3. In this case, semantic enrichment is applied to the recognized concepts. The generated question can include synonyms, hyponyms, or hypernyms (based on a particular example, a question related to a more general concept is formed). This option retains the disadvantages of the option MS.1-MC.3.
- MS.2-MC.2-MC.4. In this case, the logical models using the rules and untrained algorithms based on dictionaries, thesauri, and text corpora are applied to the original question layouts (which were generated in the subconsciousness of the system). For example, to combat typos, an algorithm based on the Levenshtein distance can be used, and rules can be used to eliminate word inconsistencies in the text.
- A complete workflow, including all modules. In this case, the hybrid question-building module uses the results of the work of the logical and machine learning modules. A unified metagraph data model [11] facilitates data integration when implementing a hybrid approach. The quality assessment module identifies the best question model based on the data generated by previous modules.

Thus, the proposed approach allows implementing a test bench that allows us to conduct the experiment with both logical methods and machine learning methods.

4. The metagraph embedding and storage structure details

In this research, the metagraph model [10] was used to describe the data structure of the text and interaction with other sources.

Graph structures are basically used to represent textual information, as they reflect the dependencies between different text parts, allow to represent external data connections, and because of that, increase the enrichment of the text. However, the description of the structure of text and knowledge in the general case in the form of a graph can be limited, since it does not allow natural hierarchies of entities to be constructed.

For experiments, it was chosen to use a simple graph structure, representing the connection between first names, last names, and names. First of all, a full list of the first names and last names was downloaded. Then the data was transformed into the graph using the new connection mark ‘in,’ which stated that the word that is in the process is inside the node for the first or last names. The structure of the graph is presented in Figure 3.

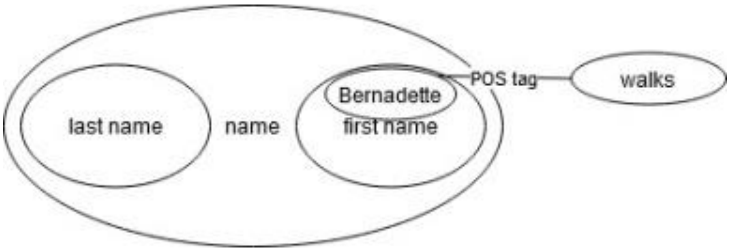

Figure 3: Example of a metagraph structure for the fully connected data

Another type of metagraph structure is presented in Figure 4. It has no connections between the first name and name and last name and name.

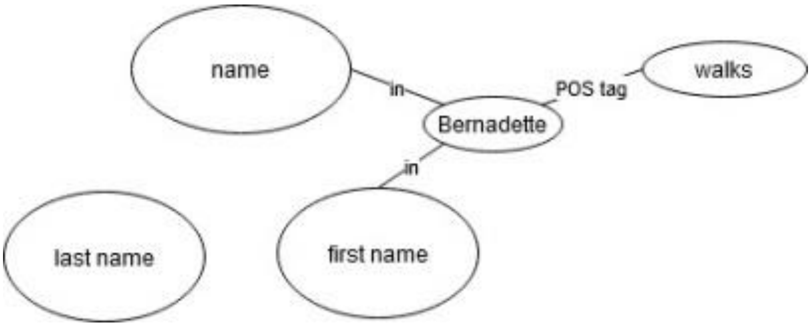

Figure 4: Example of a metagraph structure for the partially connected data

5. Experiments and evaluation

Just as in the machine translation, the most common metric for the quality of the model is BLEU [8] – the intersection of n-grams between the machine-generated text and the manually provided text.

$$BLEU = e^{\min\left(0,\frac{n-L}{n}\right)\prod_{i=1}^{N} P(i)\frac{1}{N}},$$

where n – the number of words in the reference text, L – the number of words in the output of the model, P(i) – the amount of matching i-grams for the generated and reference text, N – the maximum n for n-gram, which is, in case of BLEU-4, equal to 4.

For evaluation, the model from [5] was used, with an additional metagraph module from the section above, that encoded input data into the metagraph form before training. After the reencoding to the metagraph form, this data was converted back to fit into the graph neural network.

The structure of the neural network system is shown in Figure 5. In Table 1, the comparison between the proposed model with metagraph encoding for the fully connected metagraph structure, partially-connected metagraph structure, and the base model with no metagraph structure is presented.

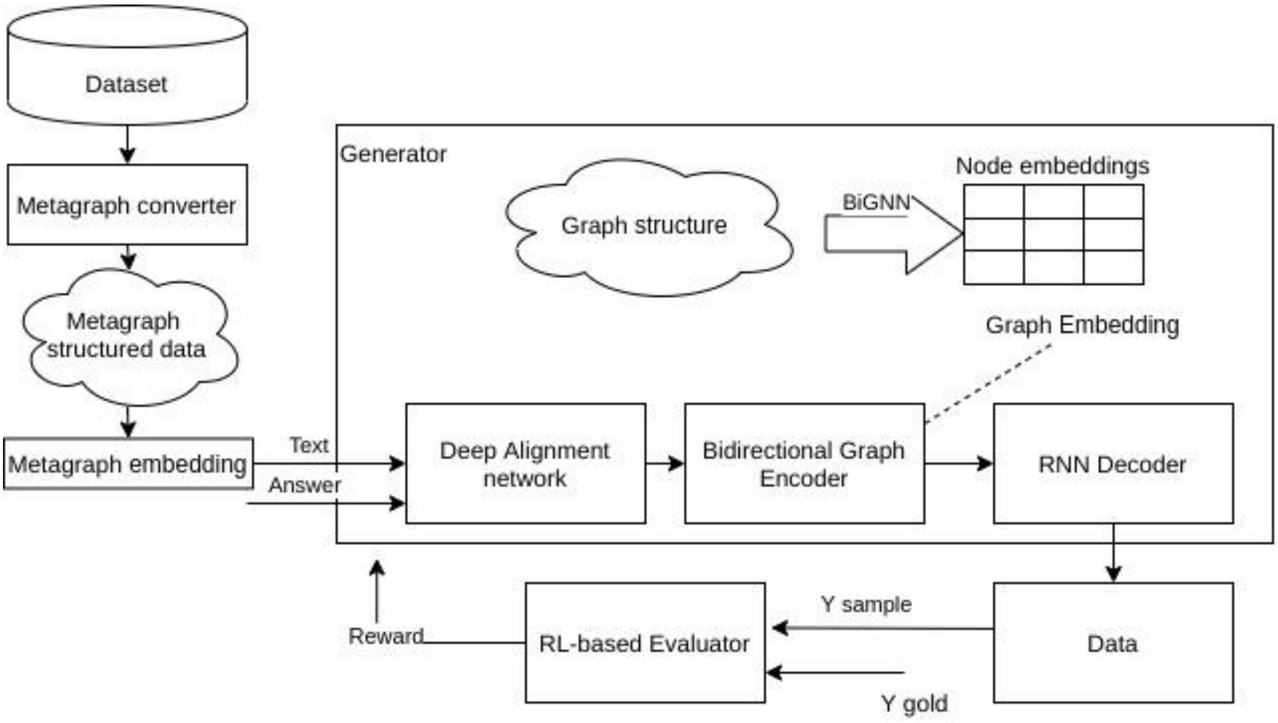

Figure 5: Structure for used neural network

All the models from the Table were trained on NVIDIA GeForce GTX 1050 graphic card with CUDA 10.2 using the SQuAD dataset split-1.

Table 1
Results of models evaluation

Name	Training duration	Quality (BLEU-4)
The model with fully connected data	788,18 mins	0.14650
The model with partially connected data	625,50 mins	0.15066
Base model	619,25 mins	0.15045

6. Conclusion

As one can see from Table 1, it takes more time to train for the model with fully connected data, probably due to the increased amount of nodes and connections between them. The quality also is lower, perhaps due to the excess amount of features used for training. The model with a partially connected data structure also takes more time to train, but the quality is better than the model with the fully connected data, and the quality is also better than the quality of the base model, but the improvement may be considered not significant.

Despite this, in this research, it was proved that the metagraph model could be used for constructing datasets for natural question generation, and in further research, the quality would be increased with the increasing complexity of metagraph edges connection and using a different thesaurus for the complex data structure.

7. References

[1] J. A. Walsh, B. D. Sattes, Quality questioning: Research-based practice to engage every learner. Corwin Press, 2016.

[2] R. Mitkov, Le An Ha, Computer-aided generation of multiple-choice tests. In: Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing - Volume 2 (HLT-NAACL-EDUC ’03). Association for Computational Linguistics, USA, 2003, 17-22, doi:10.3115/1118894.1118897.

[3] K. Stasaski, M. A. Hearst, Multiple choice question generation utilizing an ontology. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Copenhagen, Denmark, 2017, 303-312

[4] M. Heilman, N. A. Smith, Question generation via overgenerating transformations and ranking. Technical report no. CMU-LTI-09-013. Carnegie-Mellon University, Pittsburgh, 2009.

[5] Y. Chen, L. Wu, M. J. Zaki, Reinforcement learning based graph-to-sequence model for natural question generation. arXiv preprint arXiv:1908.04942 (2019)

[6] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, The Graph Neural Network Model. IEEE Transactions on Neural Networks, 20, 1 (2009) 61-80.

[7] T. Menezes, C. Roth, Semantic Hypergraphs. arXiv preprint arXiv:1908.10784, 2019.

[8] K. Papineni, S. Roukos, T. Ward, W. -J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, 311-318.

[9] V. Chernenkiy, Yu. Gapanyuk, V. Terekhov, G. Revunkov, Y. Kaganov, The hybrid intelligent information system approach as the basis for cognitive architecture. Procedia Computer Science, 145 (2018) 143-152.

[10] A. Basu, R. W. Blanning, Metagraphs and their applications. Springer, 2007

[11] V. M. Chernenkiy, Yu. E. Gapanyuk, A. N. Nardid, A. V. Gushcha, Yu. S. Fedorenko, The Hybrid Multidimensional-Ontological Data Model Based on Metagraph Approach. In: Petrenko, A., Voronkov, A. (eds.) Perspectives of System Informatics, PSI 2017, LNCS, 10742 (2018) 72-87. doi:10.1007/978-3-319-74313-4_6