Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

AN EMPIRICAL INVESTIGATION OF LANGUAGE MODEL BASED REVERSE TURING TEST AS A TOOL FOR KNOWLEDGE AND SKILLS ASSESSMENT

Tarasov D. (dtarasov3@gmail.com),
Matveeva T., **Galiullina N.**
Meanotek, Kazan, Russia

Automating assessment of person's skills is an important area of study in artificial intelligence and natural language processing. In this work we conduct empirical study of a recently proposed Reverse Turing Test for Knowledge Assessment approach—a completely automated domain agnostic method of knowledge assessment that can operate completely without human assessor involvement. Our study involved 53 participants and three different knowledge domains. We conclude that this method can reliably differentiate between expertise levels and therefore can be a compelling alternative to human grading and multiple-choice tests in many domains.

Keywords: Knowledge assesment, Reverse Turing test, language model

DOI: 10.28995/2075-7182-2020-19-696-707

ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ ОБРАТНОГО ТЕСТА ТЬЮРИНГА НА ОСНОВЕ ЯЗЫКОВОЙ МОДЕЛИ КАК ИНСТРУМЕНТА ОЦЕНКИ ЗНАНИЙ И НАВЫКОВ

Тарасов Д. (dtarasov3@gmail.com),
Матвеева Т., **Галиуллина Н.**
ООО «Меанотек», Казань, Россия

Автоматическая оценка знаний и навыков человека является важной задачей искусственного интеллекта и обработки естественного языка. В этой работе мы описываем эмпирическое исследование недавно предложенного метода обратного теста Тьюринга для оценки знаний—полностью автоматизированного метода оценки знаний, независимого от предметной области, который может использоваться без участия человека. В нашем исследовании приняли участие 53 участника и три разные области знаний. Мы пришли к выводу, что этот метод может надежно дифференцировать уровни знаний и, следовательно, может стать альтернативой человеческим оценкам и тестам с выбором вариантов ответа.

Ключевые слова: оценка знаний, тест обратного Тьюринга, языковая модель

1. Introduction

Testing student's knowledge is a cornerstone of every educational system, as no educational process can function efficiently without a way to asses that students possess certain knowledge and are capable to perform reasoning based on that knowledge.

Today, traditional ways to test knowledge, such as oral exam with a teacher, or written test, with manual verification of the results, can no longer satisfy practical needs for objective standardized and automated process.

Furthermore, the quality of human assessment of knowledge can only be as good, as knowledge of a given human assessor. As the population of the world ages, old experts leave organizations, which often leads to a catastrophic loss of critical operational knowledge [DeLong et al, 2004]. As a consequence, organizations lose not only operational efficiency, but also the ability to accurately asses knowledge of new employees and students, which opens the door to further knowledge loss.

These and other factors create high demand for computerized knowledge assessment tools and many methods has been developed to automate this process, none

of them being completely satisfactory, despite receiving wide adoption. In fact, tools such as multiple choice tests, constructed answer tests, and automatic essay grading are often considered as harmful to learning process [Ryan and Weinstein, 2009]; [Groothuis, 2018], as they favoring rote learning and can undermine both student engagement and best teaching practices. Better knowledge assessment tools are needed to solve these problems.

The Reverse Turing Test (RTT) is a variation of the Turing Test, in which the roles of computer and person are reversed [Baird et al, 2003]. In a typical implementation of the Reverse Turing Test, a computer algorithm should determine if it is dealing with a person, or with another computer algorithm. A widely known version of Reverse Turing Test is CAPTHA—a test whose task is to determine whether a website visitor is a bot or a person. CAPTHA presents the agent being evaluated with a task specially selected in such a way that it is easy for a person to solve, but is practically unsolvable for the algorithm. Many variations of this CAPTHA test [Kochanski et al, 2002]; [Lopresti, 2005]; [McInerny et al, 2019] are known.

In the context of medical diagnostics, Reverse Turing Test have been proposed for diagnosing cognitive disorders such as autism spectrum disorders (ASD) and Alzheimer's disease [Montenegro et al, 2017]; [d'Arc et al, 2018]. In particular, the papers argue that those suffering from ASD lack "the theory of mind"—the ability to predict the internal states of another person. The authors therefore propose to detect such disorders through a competitive dyad game against a computer or human opponent. Persons suffering from cognitive impairment are not able to use knowledge about the nature of the opponent (human or computer) to adapt their strategy.

Quite recently, another interesting variation of Reverse Turing Test was proposed that can be theoretically viewed as nearly universal knowledge assessment tool [Tarasov, 2019]; [2020]. In that specific version of the Reverse Turing Test a generative domain model is used to produce an object that imitates the result of human intellectual activity and knowledge is assessed based on assessing the difference between a person's interaction with a real and algorithmically generated object.

The key idea here is that algorithmically generated object will contain certain flaws due to limitation of generative model used, and these flaws should be apparent to a human who posses specialized knowledge to identify them. Such approach circumvents the biggest barrier to automatic knowledge assessment—the lack of human expert level AI, making possible for lesser intelligence to assess knowledge of the higher one.

One practical implementation of such test is language model-based reverse Turing test (LM-RTT). In this implementation a high-capacity neural language model is trained on a set of texts containing specialized knowledge, such as scientific publications on a given subject. An examinee is then given a mix of generated and real text fragments from with the the task of distinguishing real and fake texts. The hypothesis is that since modern language models are known to be capable of generating text, that can not be reliably identified as fake by humans using only grammatical cues and common sense [Graefe et al, 2018], passing such a test will require examinee to catch world-modeling failures in their respective fields of study. The task therefore, will require construction of the mental model of certain situation and checking it for consistency, and be impossible to complete with just rote learning.

Can such a procedure really be a reliable indicator of student's knowledge? A positive answer to that question can lead to development of a completely new knowledge assesment method, possesing unique properties of being fully automated and capable to asses deep knowledge and reasoning abilities.

To answer this question, we conducted a number of experiments where we compare LM-RTT scores of established domain experts in certain subject area versus non-experts, or students. Our results indicate that experts score significantly higher, and that it is possible to distinguish member of expert from non-expert group with high accuracy.

2. Related work

A number of attempts to automate knowledge assessment were made. A multiple-choice test (MTT) and its numerous variations received a wide adoption after early computer systems allowed for automated scoring of such tests. They, however, have a well-known disadvantages. MTT and its variants favor rote learning, can not assess higher-order reasoning skills and still need human to develop it. They are also prone to ambiguous interpretation problem [Ryan et al, 1998]; [Roediger, 2015].

There are many known approaches to the computer generation of new MTTs using the knowledge base (ontology) [Papasalouros et al, 2008], as well as more modern solutions allowing to rephrase key sentences of a text into a question using semantic analysis [Kantor et al, 2018] or deep neural networks [Subramanian et al, 2017] and the generation of wrong answers. However, automatically generated MTTs are prone to errors, usually require human supervision and even less suited to asses higher-order reasoning and thinking capacity then human-authored tests.

Another branch of research aims to apply natural language understanding for verification of free-form answers. Automatic essay grading [Dong et al, 2017], theme adherence check [Tikhomirov et al, 2019], and plagiarism detection [Zubarev et al, 2019] are examples of such attempts. However reported performance even for relatively simple tasks like topic adherence is low, and checking higher-order reasoning in free-form answers outside of a few very specialized cases remains currently out of reach for present day technology. In fact, it can be argued that these forms of knowledge assessment require AI to posses human or even super-human level of general intelligence, since assessing knowledge of an agent can only be done by agent that possess same or better level of understanding of the subject.

3. Algorithms and Methods

3.1. Test groups

The lack of gold standard knowledge assessing method prevents us from directly comparing gold standard student grades with RTT grades. Lack of correlation between RTT and human grades or multiple-choice test scores can result from both invalidity of LM-RTT or failure of commonly used methods to capture real knowledge

score. We opt therefore for a more objective approach. In our study we compare score of a group of students or non-experts with the score of a group of experts where experts are selected on the basis of a) having real working experience in the field of interest for at least 2 years b) being active in the in the field of interest at the moment of the study. A good method of knowledge assessment should be able to differentiate between students group and experts group.

All participants were recruited on condition of anonymity and gave permission to publish aggregate results statistics. However, we were not able to obtain permissions from corresponding institutions to mention these institutions names, therefore institutions are described here in generic terms.

Due to preliminary nature of this study, the total number of participants was relatively small. We believe that this is justified because preliminary evidence of method validity is need to justify expensive and labour-intensive larger studies.

3.1.1. Computer science domain

The topics of the test were "HTML and OSI model". Main group consistent of 28 students (last year of study) of information technology college. Comparison group consisted of 5 software developers with 2 years of work experience.

3.1.2. Biomedical domain

For this domain we compare scores of 3 groups. Main group consisted of 7 students of department of biochemistry at a major university. Comparison group #1 consisted from 5 lecturers in the same department, including two PhDs. Comparison group #2 (non-biomedical experts) consisted of 5 software developers (the same as previously). The aim for having comparison group #2 was to asses whenever RTT measures specialized knowledge or just the level of general intelligence or ability to identify machine-generated text by looking for specific flaws. The topic of the test was "Proteins secondary and tertiary structure, structure of collagen and insulin".

3.1.3. Food safety domain

Unlike two previous groups this test measured the knowledge of internal document (Restaurants rules for personnel). Main group consisted of 3 new hires, comparison group #2 consisted of 3 managers with 4 years+ experience, comparison group #3 consisted of 2 software developers

3.2. Language models and test construction

3.2.1. Computer science domain

16 paragraphs were sampled from English Wikipedia pages on HTML and OSI model. Paragraphs were required to contain keywords "HTML" or "OSI" in first 10 words. After that 16 paragraphs were generated with GPT-2 large model [Radford et al, 2019] using top k random sampling with $k = 2$, using first 10 words as context for generation (to make model follow desired topic). All texts were then translated into Russian using Google Translate API. Such approach makes it impossible for students to find real paragraphs by searching Internet.

A complete test, containing with 32 text fragments and the task was to distinguish between real and generated fragments on the basis of their logical consistency and factual correctness.

3.2.2. Biomedical domain

For biomedical domain, custom language model was trained on corpus of freely available pubmed abstracts (https://www.nlm.nih.gov/databases/download/pubmed_medline.html), data from 2019 baseline were used. We used character-level LSTM-based model [Hochreiter and Schmidhuber, 1997] with 7 LSTM layers, with 3,192 LSTM units in first two layers and 2,500 units in remaining layers. The model was trained for 1 month using two 2080Ti GPUs, achieving 0.96 BPC on test set.

For construction of the test, 200 abstracts were found with pubmed search API using keywords "collagen structure" and "proteins structure and function" out of which 14 abstract where sampled randomly and 14 abstracts were generated with language model using nucleus sampling method and corresponding article title as a context. We then used first three sentences from each abstract as a text fragment for construction of complete test (total 28 text fragments). All texts were translated with into Russian using Google Translate API.

3.2.3. Food safety domain

Custom language model was pre-trained on complete Russian Wikipedia dump (8Gb of text). We used character-level LSTM-based model with 8 LSTM layers, with 3,192 LSTM units in first two layers and 2,500 units in remaining layers. The model was trained for 2 weeks using two Tesla V100 16GB GPUs, achieving 1.04 BPC on test set. The model was then fine tuned on internal document, that described the food safety rules to which restaurant personnel must obey. Total document size was 260KB. We then randomly selected 14 paragraphs from this document and sampled corresponding 14 paragraphs from the model, using first 10 words of real paragraphs, as context.

3.3. Analysis of results

Statistical analysis was conducted by means of exact binomial test. The null hypothesis was that probability of giving a correct answer is the same in all groups.

4. Results and discussion

4.1. Computer science domain

For students group, average number of correct answers was 16.1 (50.3%), maximum 23 correct answers, minimum—11 correct answers. For experts group, average number of correct answers was 21.2 (66.25%), maximum was 24, minimum—18. Threshold of 21 correct answers allowed to distinguish student from expert with 98% accuracy with one false positive (student identified as expert) and one false negative (expert identified as student). The differences between means in two groups were found to be statistically significant at $p=0.95$ level using exact binomial test.

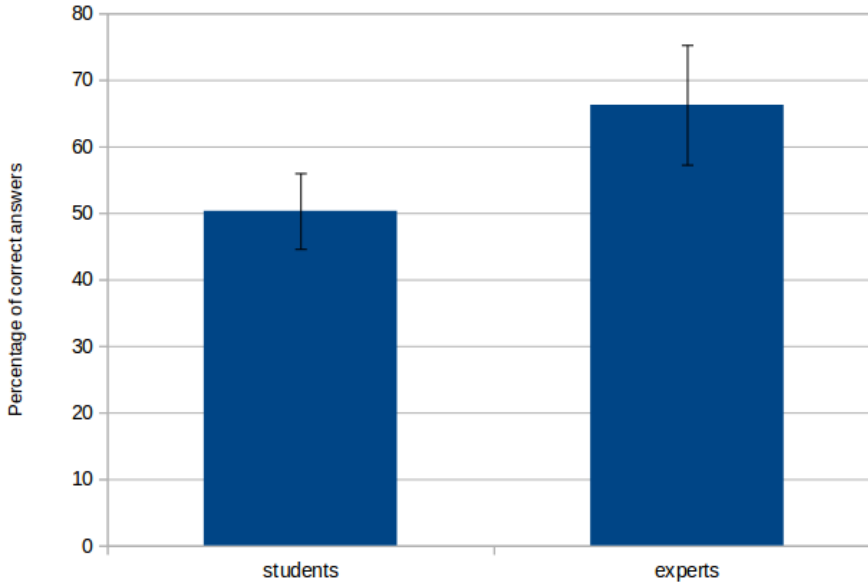

Figure 1. Percentage of correct answers in students and experts groups (Computer science domain). Error bars indicate 95% binomial confidence interval

No correlation was found between RTT scores and student's annual grades. In fact, student with highest mark RTT (23) had the smallest annual grade (3). The teacher explained his grade was for low attendance and poor behavior. This leaves out 3 possible explanations:

- Administered RTT test was not sensitive enough to distinguish between students knowledge
- Test scores reflected higher general intelligence in expert group, not specialized knowledge
- Human annual grades were subjective, not reflecting actual knowledge

4.2. Biomedical domain

Average score was 52% for student group, 67.5% experts group and 56.2% in no biomedical expertise group. The difference was statistically significant at p=0.95 level between student and expert groups and between no biomedical expertise and expert group, using exact binomial test. These results suggest that specialized knowledge gives advantage to biomedical experts over both students and software developers groups, even though software developers group was previosly exposed to LM-RTT test on different topic and generally familiar with artifiacts of neural text generation.

Another interesting point is that software developers average score on in-domain test, described above was significantly higher then on biomedical test. Taken together these results indicate that LM-RTT score is improved when examinee posseses specialized knowledge. These results also indicate that humans can't easily adapt to solve such tests without having necessary knowledge.

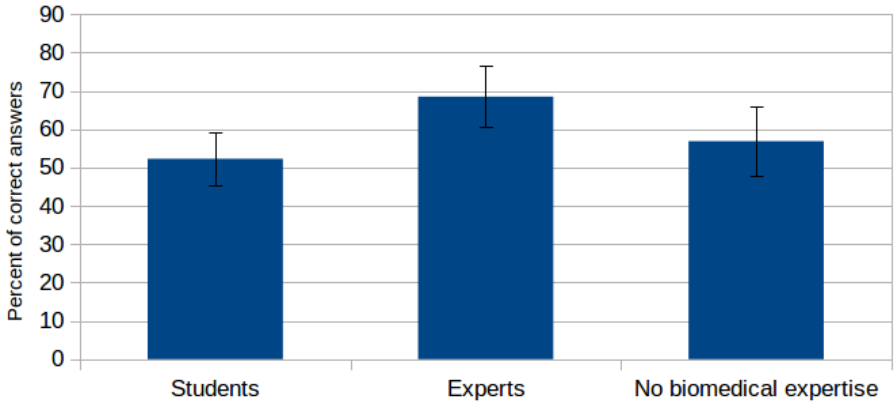

Figure 2. Percentage of correct answers in students and experts groups and group without biomedical expertise (Biomedical domain). Error bars indicate 95% binomial confidence interval

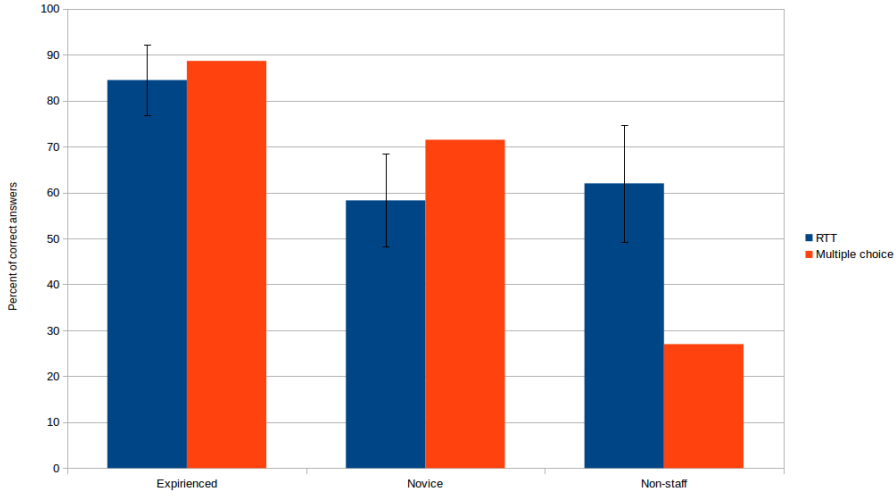

Figure 3. Percentage of correct answers in experienced, novice staff groups and non-staff group (Food safety domain) for RTT score and handcrafted multiple-choice test score (regularly used by restaurant to assess personnel knowledge). Error bars indicate 95% binomial confidence interval

4.3. Food safety domain

For experienced personnel group, average number of correct answers was 23.3 (84.5%), maximum 24 correct answers, minimum—23 correct answers. For novice group, average number of correct answers was 16 (57%), maximum was 18, minimum—14. The differences between means in experienced and novice groups were found to be statistically significant at p=0.95 level using exact binomial test. There were no significant difference between novice and non-staff groups.

The unexpected finding here was that while standard multiple-choice test, regularly used by restaurant to assess personnel knowledge, demonstrated clear difference between novice and non-staff groups, there were no such differences using RTT. At the same time, both tests are showing difference between experienced and novice. Restaurants rules require all new staff to pass multiple choice test and thus people repeatedly attempt to pass it and test content is not being changed between attempts, thus people in novice group could learn correct answers but their actual knowledge is low. In contrast, experienced staff group possesses real knowledge about operational procedures.

4.4. Qualitative analysis

Table 1 shows some example of generated fragments in two domains and shows several types of typical errors.

Table 1. Examples of generated fragments

Domain	Text	Error type
Computer science	HTML markup consists of several key components, commonly known as the head element, the body element, and the script element. Body element is often referred to as the document root (or document head), as that is where HTML documents are created. The body element contains information about the document's contents, such as the document title and document meta data, which are not normally presented elsewhere on the page.	World modelling failure: HTML body is described as a place where HTML documents are created. Factual error: document metadata are located in document head, not body.
	HTML documents imply a structure of nested HTML elements. This is called a tree. If each node contains an attribute and some value, this might look like this: <html> <head> <title>hibernate generator example</title> </head> <body> test </body> </html> A node in this tree has an attribute named "title", and a value of "test".	Logical consistency error: element "title" does not have value of "test" in this example

Domain	Text	Error type
Biomedical	The aim of this study was to investigate the effect of BMP-2 on osteoclast formation and osteoclast formation in vitro.	Neural text degeneration (repeated phrase “osteoclast formation and osteoclast formation in vitro”)
	Collagen is the most abundant protein in the blood serum of patients with chronic renal diseases.	Factual error—collagen is not most abundant protein in blood serum, serum albumin is.
	The aim of this study is to identify, in terms of the collagen content, the proteoglycan content of the collagen fibrils in the developing human and rabbit condylar cartilage. The collagen fibrils were isolated from bovine cartilage by differential centrifugation.	Logical inconsistency. First sentence speaks about human and rabbit subjects, while second tells that collagen fibrils were isolated from bovine cartilage
	The aim of this study was to evaluate the effect of collagen type I on the production of collagen in rat skin by measuring the changes in the cell content of the serum proteins and the activity of the enzyme collagenase. Twenty-four hours after the induction of chronic progressive renal failure, the collagen content of the renal cortex was measured by radioimmunoassay. The results showed that the protein content of the collagen fibrils was significantly increased in the regenerated renal tissue.	World model error: collagen production is rat skin can not be measured by changes in «cell content of the serum proteins». Logical inconsistency (first part talks about collagen in skin, and then about collagen in renal cortex). Described overall sequence of events is biologically implausible. Grammar and overall structure, however are realistic.

We found a mix of surface mistakes, that stem from neural text degeneration, logical inconsistencies, factual errors and complex world modelling errors. Clearly, there is a room for improving test quality by using better language models and taking measures to prevent samples with degenerated text from appearing in the test.

5. Conclusions

We have established that:

1. Language model based reverse Turing test can reliably distinguish experts from non-experts in all 3 domains studied in this paper.
2. We found that specialized professional knowledge result in higher LM-RTT scores then using grammatical and logical consistency cues alone.
3. We repeatedly observed certain discrepancies between LM-RTT and traditional knowledge assessment methods, such as human grading and multiple choice test. We can’t confidently establish reason for these discrepancies.
4. No significant adaptation was observed due to repeated exposure to LM-RTT testing format.

Overall LM-RTT seems to be a promising method for knowledge assessment. Undoubtly, due to novelty of the method and controversial nature of RTT hypothesis, a lot of large studies in different knowledge domains will be needed to establish

method validity with enough certanity. However, such studies are difficult and costly to conduct as they need to involve a lot of real human subjects and some prior evidence is required to justify these expenses, and our findings justify further investigations of method capabilities using larger groups.

We believe that our results are important, because, if confirmed by further studies, they can lead to development of realively cheap but powerful automated method to asses depth of person's knowledge in many domains.

References

1. *Baird, H. S., Coates, A. L., & Fateman, R. J.* (2003). Pessimalprint: a reverse turing test. International Journal on Document Analysis and Recognition, 5(2–3), 158–163.
2. *d'Arc, Baudouin Forgeot, Devaine Marie, and Jean Daunizeau.* "A reverse Turing-test for predicting social deficits in people with Autism." bioRxiv (2018): 414540.
3. *DeLong, David W., and J. Storey.* Lost knowledge: Confronting the threat of an aging workforce. Oxford University Press, 2004.
4. *Dong, Fei, Yue Zhang, and Jie Yang.* "Attention-based recurrent convolutional neural network for automatic essay scoring." Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). 2017.
5. *Graefe, A., Haim, M., Haarmann, B., & Brosius, H. B.* (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. Journalism, 19(5), 595–610.
6. *Groothuis, Douglas.* "How Multiple-Choice Tests and Machine-Graded Essays Undermine Learning." Academic Questions 31.1 (2018): 70–76.
7. *Hochreiter, Sepp, and Jürgen Schmidhuber.* "Long short-term memory." Neural computation 9.8 (1997): 1735–1780.
8. *Kantor, Arthur, Jan Kleindienst, and Martin Schmid.* "Automatic question generation from natural text." U.S. Patent No. 9,904,675. 27 Feb. 2018
9. *Kochanski, Greg, Daniel Lopresti, and Chilin Shih.* "A reverse turing test using speech." Seventh International Conference on Spoken Language Processing. 2002.
10. *Lopresti, D.* (2005, May). Leveraging the CAPTCHA problem. In International Workshop on Human Interactive Proofs (pp. 97–110). Springer, Berlin, Heidelberg.
11. *McInerny, Michael James, Mark Evans Brighton, Sevag Demirjian, and Blair Livingstone Hotchkies.* "Turing test via failure." U.S. Patent 10,262,121, issued April 16, 2019.
12. *Montenegro, Juan Manuel Fernandez, and Vasileios Argyriou.* "Cognitive evaluation for the diagnosis of Alzheimer's disease based on turing test and virtual environments." Physiology & behavior 173 (2017): 42–51.
13. *Papasalouros, A., Kanaris, K., & Kotis, K.* (2008, July). Automatic Generation of Multiple Choice Questions From Domain Ontologies. In e-Learning (pp. 427–434)
14. *Radford, Alec, et al.* "Language models are unsupervised multitask learners." OpenAI Blog 1.8 (2019): 9.
15. *Roediger III, Henry L., and Elizabeth J. Marsh.* "The positive and negative consequences of multiple-choice testing." Journal of Experimental Psychology: Learning, Memory, and Cognition 31.5 (2005): 1155

16. *Ryan, Ann Marie, and Gary J. Greguras.* "Life is not multiple choice: Reactions to the alternatives." Beyond multiple choice: Evaluating alternatives to traditional testing for selection (1998): 183–202.
17. *Ryan, Richard M., and Netta Weinstein.* "Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing." Theory and research in education 7.2 (2009): 224–233.
18. *Subramanian, Sandeep, et al.* "Neural models for key phrase detection and question generation." arXiv preprint arXiv:1706.04560 (2017).
19. *Tarasov D. S.* A method for creation of tests. Patent application (Russia) N 2020103928 from 29.01.2020
20. *Tarasov D. S.* A way to create test items to test the depth of knowledge and the ability to reason students and specialists. Patent application (Russia) N 2019127875 from 04.09.2019
21. *Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V.* (2019). Assessing Theme Adherence in Student Thesis. Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference "Dialogue", Issue 18
22. *Zubarev D. V., Sochenkov I. V.* (2019). Cross-Language Text Alignment for Plagiarism Detection Based on Contextual and Context-Free Models. Computational Linguistics and Intellectual Technologies: Proceedings of Annual International Conference "Dialogue", Issue 18.