**УДК 004.855**

# ON USAGE OF MACHINE LEARNING FOR NATURAL LANGUAGE PROCESSING TASKS AS ILLUSTRATED BY EDUCATIONAL CONTENT MINING

**A.V. Melnikov[1], D.S. Botov[2], J.D. Klenin[3]**

*Chelyabinsk State University, Institute of Information Technology, Chelyabinsk, Russia*
*mav@csu.ru[1] , dmbotov@gmail.com[2] , jklen@yandex.ru[3]*

**Abstract**

In this paper, we review most popular approaches to a variety of natural language processing (NLP) tasks, primarily those, which involve machine learning: from classics to state-of-the-art technologies. Most modern approaches can be separated into three rough categories: ones based on distributional hypothesis, those extracting information from graph-like structures (such as ontologies) and the ones that look for lexico-syntactic patterns in text documents. We focus mainly on the former of the three. Before the analysis can even begin, one of the important steps in preparation stage of NLP is the task of representing words and documents as numeric vectors. There exists a variety of approaches from the most simplistic Bag-of-Words to sophisticated machine learning methods, such as word embedding. Today, in the task of information retrieval the best quality for both English and Russian languages is achieved by approaches based on word embedding algorithms, trained on carefully picked text corpora in conjunction with deep syntactic and semantic analysis using various deep neural networks. A big variety of different machine learning algorithms is being applied for NLP tasks such as Part-of-Speech-tagging, text summarization, named entity recognition, document classification, topic and relation extraction and natural language question answering. We also review possibilities of applying these approaches and methods to educational content analysis, and propose the novel approach to utilizing NLP and machine learning capabilities in analyzing and synthesizing educational content in a form of a decision support systems.

***Key words:*** *machine learning, natural language processing, educational data mining, semantic similarity, deep learning, neural networks.*

***Citation:*** *Melnikov AV, Botov DS, Klenin JD*. On usage of machine learning for natural language processing tasks as illustrated by educational content mining. *Ontology of designing*. 2017; 7(1): 34-47. DOI: 10.18287/2223-9537-2017-7-1-34-47.

## Introduction

Development of methods of intelligent text analysis is one of the key problems in the field of Artificial Intelligence (AI) research. Tasks, related to this problem, are usually referred to as Natural Language Processing (NLP). It's an interdisciplinary area of science and technology, aimed to resolve the problems of automatic analysis and synthesis of natural language that appear during man-machine interactions, using various AI and computer linguistics approaches.

Until recently, despite scientists' best efforts accuracy and recall of such methods couldn't possibly compare to results, demonstrated by a human. At best, results that were worth speaking of, were achieved only for a limited area of knowledge or a select range of text properties. However, development of machine learning techniques made it possible to achieve quality, required for practical use in NLP tasks. Nowadays, it is possible to propose a problem of deep information extraction from text to be used in creation of formal models of specific areas of knowledge.

## 1 Main tasks and approaches to natural language analysis

Most common tasks in NLP are:

- semantic similarity and relatedness evaluation;
- information retrieval;
- information extraction (named entity recognition, relation extraction, fact extraction, knowledge extraction, coreference resolution);
- text classification and text clustering;
- natural language question answering;
- machine translation;
- text summarization;
- sentiment analysis and opinion mining;
- automated ontology/dictionary/thesaurus/knowledge base generation;
- speech recognition and speech synthesis;

Before approaching more complex and specific tasks, it is important to find a representation for a text, through the use of text models (See Figure 1).

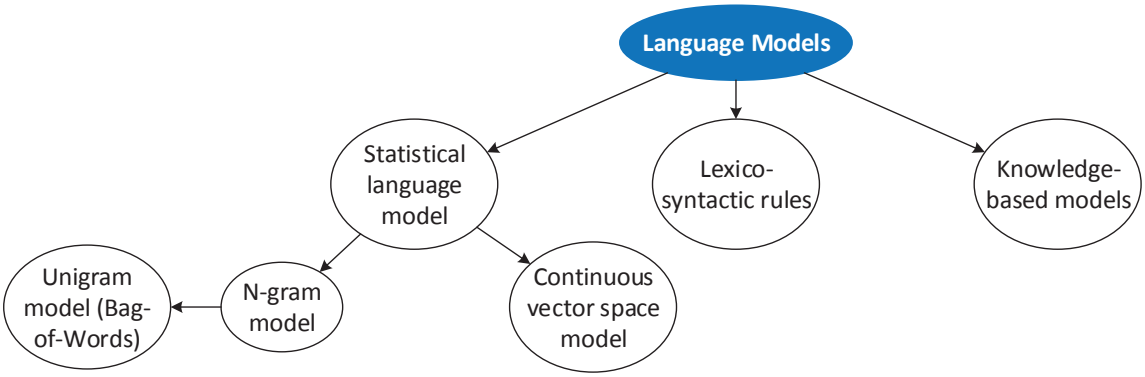


Figure 1 – Language modelling approaches

The most simplistic approach to language modeling includes various statistical models, based primarily on word distributions within the document or a collection of documents i.e. distributional semantics.

Distributional semantics approach defines semantic similarity between two linguistic items (such as words or words combinations) based on their distributional properties in large corpora of text, generally without specific knowledge of lexical or grammatical meanings of said items. The main idea behind this approach is so-called distributional hypothesis: linguistic items with similar distributions have similar meanings. Various vector models and word embedding techniques that transform each text item (words, usually) into numeric vectors are commonly used in context of distributional semantics.

One of the ways to represent words with those ideas in mind is to cut documents into sets and sequences of words – n-grams [1] or shingles, which account for information contained within multi-word constructs of length n: bigrams for word pairs, trigrams for triples and so on.

The edge case scenario for n-gram model occurs with n equal to 1. Such model could be called unigram, but more often referred to as the Bag-of-Words model [2]. Such model disregards all document properties except for the counts of words in it. Bag-of-Words represents a set of documents as a matrix with rows corresponding to documents and columns – to a specific term. Values on the intersections describe the count of a word in a specific document.

As multi-word constructs contain additional information, when compared to a set of singular words (word collocations, idioms and so on), it is an important step in text preparation to distinguish bigrams and trigrams to keep, while breaking the rest of the document into a Bag-of-Words.

In practice, BoW models usually include some sort of weight for each term-document pair. The simplest and most obvious measure would be the number of occurrences (frequency) of a term in each document within collection, or normalized probability of finding the word within the document. This, however, evaluates more common words as more important, while that might not necessarily be true. The most common weighting measure for both singular words and n-grams, which counteracts this bias is TF-IDF (TF — term frequency, IDF — inverse document frequency) [3, 4]. It is a statistic, used to determine the importance of the word in the context of a document, which itself is part of a document collection or a corpora. The weight of the word in a certain document is proportional to its count in said document and at the same time inversely proportional to this word's frequency in other documents from the same collection.

One of the early methods utilizing distributional semantics was Latent Semantic Analysis (LSA). This method defines a vector for each word based on results of applying singular value decomposition (SVD) to a weighted term-document matrix, containing word counts. All word-vectors together represent vector space.

One of the main problems for such models is a so-called "semantic gap", which results in sparse vectors and is caused by ratio of unique words existing in the language to the count of unique words that appear in a single document.

Distributional semantics approaches are attempting to solve the task of sparse word vector space dimensionality reduction in order to reduce the effects of semantic gap.

Another statistical model – continuous space language model – helps treating semantic gap issue. This model represents text as a continuous stream of words, perceived through the context window, which includes immediate context of each word. To calculate the vector representation of each word, this type of model utilizes various word embedding techniques. Previously discussed models represent words from document collection in a form of a sparse vector space with dimensionality of entire word count of the language (measuring around one million or more for English, for example). Mathematical embedding involves transforming those vectors into a much less dimensional vector space with much denser vectors. This is usually done through various machine learning approaches, like neural networks and log-bilinear regression. The topic of word embedding is discussed further in part 4 of these paper.

A different approach to language modeling involves using a priori knowledge of lexical and syntax rules for a specific language in order to extract the inner structure of the text. Lexico-syntactic knowledge helps determining types of entities and relationships between them, based on exact words, their forms, part of speech and part of a sentence.

Many methods use lexico-syntactic patterns and syntactic-semantic analyzers as parts of the solutions to the task of information extraction. One of the possible approaches to formulate lexico-syntactic pattern is to apply context-free grammar and keyword dictionaries (Tomita-parser from Yandex, for instance), meta-languages describing rules (such as RCO Fact Extractor). PatternSim [5] is an example of a tool that uses lexico-syntactic patterns to measure semantic similarity between words.

Another approach has demonstrated its efficiency in analysis of texts in Russian. It is based on relationally-situational text model [6], communicative grammar and heterogeneous semantic networks theories. Relationally-situational methods combine static and dynamic approaches to text processing. This approach also uses dictionaries, thesauruses, ontologies and linguistic knowledge bases.

Third popular approach to text modeling involves usage of various ontologies and knowledge bases, containing specific terms, entities and their relationships. This approach is also known as graph-based or network-based approach. Due to how easy it is for a human mind to represent knowledge space as a set of objects and their interconnections, methods utilizing various graphs as a

knowledge base structure are quite popular. Such structures include semantic nets, concept maps, ontologies, thesauruses. Objects, concepts and ideas are usually represented by nodes in a graph and their connections to each other – by graph's edges.

There are many methods suggested for determining semantic similarity between concepts that are based on paths and depths of objects in a graph, such as Resnik, Lin, Jiang and Conrath, Wu and so on. Mentions in the document can be resolved into certain nodes of the underlying knowledge base, which provides an opportunity to project graph structure onto the text and discover how specific relations are represented in text. This could be used to extract syntactic patterns, which could be used to further populate the database by extracting new concept pairs with known types of relations.

Graph-like structures can be used to solve complex problems of semantic analysis. Historically local ontologies of specific fields, semantic net of WordNet, community encyclopedia Wikipedia, and other dictionaries and thesauruses were used as such structures.

Thorough overview of existing modern semantic similarity measures is presented in [7].

## 2 Machine learning in NLP: algorithms, tools and technologies

Before we introduce some of the more popular and specific approaches, it is important to understand the general structure of machine learning as a field. Traditionally specified approaches to machine learning (ML) are supervised and unsupervised machine learning. Supervised ML is based on the idea of an algorithm learning how to perform specific action on a training data set, which already has correct answers to the learning problem, while unsupervised approaches attempt to generate the answer themselves, by working directly with unlabeled data. Obviously, the main difference in practice is the training data requirement. For supervised algorithms this dataset needs to be assembled and labelled with correct answers (class tags, for example) and only then it can be used to train algorithms. This usually involves high amounts of manual work, which is unnecessary for unsupervised algorithms. Furthermore, way more data, text data in particular, is available in an unlabeled form, making unsupervised training especially interesting for NLP tasks.

As a compromise between two approaches, the semi-supervised ML is specified. Technically, semi-supervised ML is still supervised, but this umbrella term covers techniques and methods that start by using only a tiny amount of labelled data and go on from there attempting to label unlabeled data on their own, learning from the mixed dataset of labelled and unlabeled data. This approach is quite popular among researchers in the field of NLP and is used for many of the examples below.

The development of machine learning approaches has caused many scientific researchers to apply those methods to NLP tasks. In an attempt to structure popular in NLP algorithms and methods, we combined them into the rough diagram presented in figure 2.

Conditional Random Fields (CRF) classifiers are one of the popular ML algorithms in text analysis, since they can take into account not only singular words, but their context as well. CRF are used, for instance, in the task of named entity recognition in documents from various fields and areas of knowledge [8-10] and text summarization problems [11, 12] in which this classifier is applied to distinguish more important sentences.

Logistic regression algorithms were applied to named entity recognition as well, for instance in [13], in which they were applied to the problem of classification Wikipedia articles according to the types of concepts they represent. This classifier was also used to determine similarity between nodes, extracted from the graph of the knowledge base, and the correct answer as part of the natural language question-answering systems [14].

Support Vector Machines (SVM) were applied to a variety of text analysis-related tasks. In particular, in named entity recognition, SVM were used to increase quality of found entities as a last

level of a three-level framework, proposed in [15]. Another team of researchers compared SVM to CRF (with CRF showing slightly better results in majority of tests) in the task of word classification as a part of a named entities in sentences from medical documents [16] in BIO (Beginning-Inside-Outside) format.

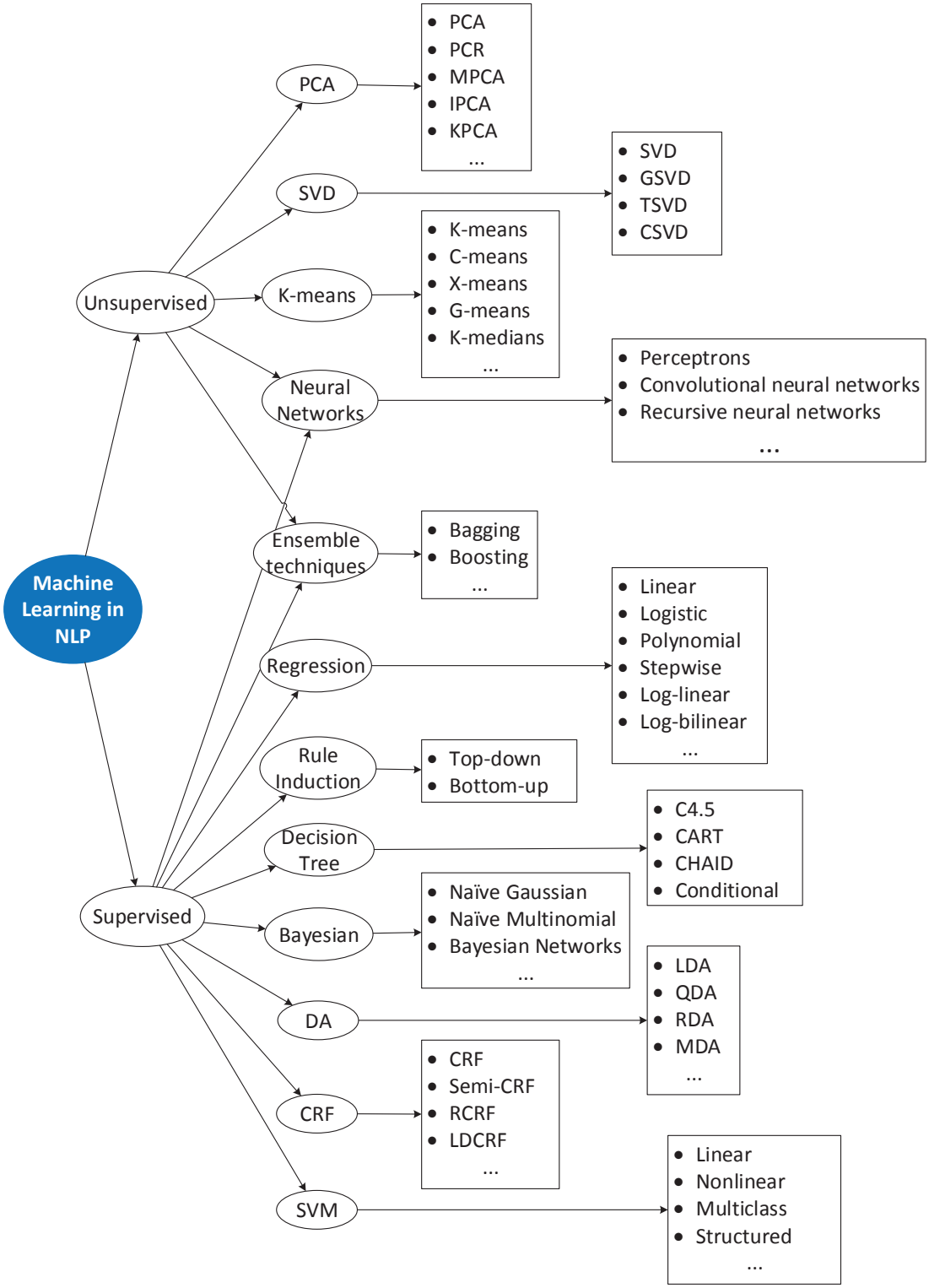


Figure 2 – Rough diagram of popular machine learning algorithms to natural language processing tasks

Among other algorithms, one could mention LDA-based LabeledLDA, applied to named entity recognition [17], as well as naive Bayes classifier and decision-tree generator C4.5, which were applied to the task of authorship recognition (classification problem) based on syntactic n-grams [18], but lost in quality to SVM.

Some of the popular information retrieval approaches are based on rules that describe certain properties of text elements. With these methods, ML can be applied to automatize the process of rule induction. One type of such systems - top-down - is based on the process of generating new, more detailed rules on the intersections of more general ones. For example, LemmaGen [19] is an part-of-speech tagger, which is based on iterative process of RDR (Ripple Down Rules).

The opposite approach - bottom-up - proposes rule induction for more general rules from specific cases and mentions of concepts in the document. These rules were used, for instance, to retrieve information about medical events from clinical records. [20].

Quite popular are ensemble algorithms, such as random forest, used, for example to extract information about medical drug interactions [21] and determining personal characteristics of a writer [22]. There was also proposed a variation of the algorithm, which was stable against imbalanced classes [23].

## 3 Machine learning approaches based on neural networks

Nowadays, approaches that utilize neural networks are rapidly grow in popularity. For instance, averaged perceptron was used for coreference resolution by the developers of Reconcile [24], as a classifier, determining the coreference probability of two noun phrases.

In natural language question-answering task a hidden variable perceptron was used to filter incorrect relations, extracted from the knowledge base [25].

Striking example of applying neural network in the context of distributional semantics is Word2Vec model, created by a team of researchers from Google, led by Tomas Mikolov [26]. Word2Vec solves the problem of generating a statistical model of natural language via analysis of large text corpora.

This model is based on earlier research [27, 28] into statistical modelling of natural language through word embedding, accomplished by using neural networks. Main idea behind this approach is to use two-layer neural network to transform text corpora into its vector form in n-dimensional vector space (with n usually being around several hundreds), based on the distribution of the original linguistic items in the corpora. The notable property of resulting vectors is that similarity between vectors reflects semantic similarity or relatedness between original linguistic items, with model being trained specifically to improve the quality of vector representation: the more distributionally similar original items are, the more similar their vector representations must be.

Word2Vec uses two models for training: continuous bag-of-words (CBOW) model, attempting to predict the word, knowing its context (word order within context window is of lesser importance) and continuous skip-gram model, which does the opposite, by predicting word's context from a single word. Continuous skip-gram shows better results for corpora that include rare words and in general achieves better quality than CBOW given a larger datasets, however it takes longer to train and is worse for determining syntactic trends.

Word2Vec on its own shows good results in calculating semantic similarity of words, determining the missing word from context, and achieves satisfying in multi-word structures' analysis. However, Word2Vec is presented by its authors as a first stage of text analysis. To solve complex problems additional methods and metrics based on specific area of knowledge are required. Semantic vectors of documents or parts can be used for further analysis, for example classification and clustering with more generic algorithms of ML (SVM, decision-trees, neural network and so on).

One of the Word2Vec's analogues is GloVe, developed by a group of researchers from Stanford [29]. It is based on global log-bilinear regression model, which combines pros of both global matrix decomposition and local context window. GloVe allows for highly efficient solutions to named entity recognition, semantic similarity tasks. For particular test sets it consistently outperforms Word2Vec, however, according to researches: exact comparison is difficult due to vast amounts of parameters in both models.

Efficiency of word embedding approach can be judged from results of Dialogue Evaluation competition. In this competition, various algorithms tested against each other in a set of text analysis tasks for Russian. During 2015s competition for semantic similarity (RUSSE) [30], majority of top-performing models were based on Word2Vec (or its analogues), trained on various Russian corpora. Other methods that proved effective when combined with Word2Vec were: decision trees over n-gram model, logistic regression and taking into account morphological properties of linguistic items.

For instance, direct comparison [31] of three systems based on lexico-syntactic patterns for syntax extraction, context window over data from Google n-grams and Word2Vec over a huge corpora, respectively shows that the best results were achieved by the latter system.

Later on the team behind Word2Vec had suggested the way to apply similar techniques to bigger textual units – sentences, paragraphs, and entire documents [32]. Similarly to Word2Vec, the new algorithm was named Paragraph2Vec, as it provides the vector space representation of a paragraph of text – paragraph vector. Overall, the algorithm is extremely similar to Word2Vec with major difference being in the utilization of a secondary matrix, containing vectors for every text or paragraph encountered during training.

Again, similarly to Word2Vec, this algorithm provides two distinctive models for training. The Distributed Memory model (PV-DM) is similar to CBOW approach in Word2Vec, using both vectors of words from the context window and the paragraph vectors to maximize the quality of predicting the missing word in its context. In this case the paragraph vector represents the topic of the entire paragraph which is missing from the immediate context of the predicted word (with this context being represented by the word vectors) – thus being the memory which is distributed across all the contexts within the paragraph.

Second model in Paragraph2Vec is called Distributed Bag-of-Words (PV-DBOW) and is similar to the skip-gram model in Word2Vec. This model uses only paragraph vectors to predict words from contexts sampled from these paragraphs. As the authors specify, PV-DM works well in most situations, but they do recommend pairing two models together to achieve a more consistent result.

## 4 Deep learning in NLP tasks

Lately, the deep learning approach became one of the breakthrough NLP technologies. Main idea behind this approach is to create models with complex structure and non-linear transformations in order to model high-level abstraction in data. The "depth" in this case is a distance in model graph between input and output nodes. In order to do so the task of training inner layers of multilayer network needs to be resolved, which can't be done by a classical ML approach of backward propagation of errors. A detailed review of deep learning structures is presented in [33]. Another extensive report, covering deep-learning approaches and algorithms and their applications in artificial intelligence was presented in [34]. That paper doesn’t focus only on those algorithms, models and techniques used and represented in NLP, but instead mainly covers semantic data mining tasks in general, as well as the uses of such algorithms in computer vision. One of the main discussed subjects of that report is knowledge bases and ontology building and how could deep learning techniques be applied to building and researching models of human knowledge.

Deep learning approach allowed for a significant improvement in speech recognition and NLP tasks, such as NLQA, sentiment analysis, information retrieval, topic modelling, text classification and text clustering, machine translation.

Most common deep learning architectures used for NLP tasks are recursive neural networks (for instance, recurrent neural networks) and convolutional neural networks.

There exists a number of works about usage of neural networks [35-38].

CNN are used for analysis of semantico-syntactic properties of the text. For example, semantico-syntactic analyzer ABBYY Compreno, shows best results for information retrieval tasks in Russian [39]. This analyzer allows one to resolve more complex tasks in information retrieval: coreference and anaphora resolution [40].

CNN as deep learning approach and SVM algorithm are often used in construction of NLQA systems. For instance, in IBM Watson CNN are used to conduct deep search of syntax patterns and answer generation [41]. In other research [42] CNN were used to transform entities and patterns of their relations in questions into vector space to be compared to already known concepts and their relations in the knowledge base.

In [43] CNN were successfully used to model natural language sentence structure.

CNN and RNN display high quality in sentiment analysis in short texts [44] and opinion mining tasks [45].

In SentiRuEval 2015 and 2016 - the Russian language sentiment analysis competition [46], best quality were achieved by systems, using Word2Vec trained on special corpora of lexicons and short documents, as a first and SVM or neural network classifier as a second stage of analysis. Among them the best results were shown by RNN and or CNN with addition of syntactic attributes to Word2Vec vectors [47].

## 5 Machine learning approaches in educational content analysis

One of the possible fields of knowledge for data analysis is educational content. The part of data mining concerned with this field is called educational data mining (EDM). Next, we'll provide a general overview of several researches of ML algorithms and semantic analysis application to the EDM tasks.

For the task of clustering of computer-supported collaborative learning participants [48] researchers used naive Bayes classifier. Another team of researchers [49] used HMM and SVM classifiers for sentiment analysis of reviews, left by users of e-learning systems, left in blogs and forums. Logistic regression model was utilized for the analysis of text recognized from speech by automatic tutor [50]. For e-learning FAQ generation [51] hierarchical classifiers and rough set theory were used. For information extraction from educational content, researchers suggested automatic framework [52] for knowledge base concept hierarchy generation from found e-learning documents on specific topic, which utilized naive Bayes classifier.

A variety of text ranking algorithms and systems designed to rate documents involved in the educational process, such as student essays, was proposed. For instance, Writing Pal [53] is a system, trained on a corpora of essays, rated by a group of experts, to predict human rating of an essay, based on wide range of linguistic, rhetorical, and contextual features as predictors in a process of stepwise regression. Another team of researchers [54] focused on developing a system capable of ranking the readability of text based on multilevel linguistic features – as in features from word, semantics, syntax and cohesion levels. Their research was performed for a corpora of text produced from Chinese textbooks and rated by a group of experts – teachers, educational psychologists and language professors. The readability is then evaluated by applying a classifier to the sets of extracted features – discriminant analysis and support vector machines.

It is important to understand, that educational content varies greatly. First of all, representation of content varies between different kinds of educational content. Secondly, even within the same educational content type, representations may still vary for documents made by different educational organizations, or even between different departments or educators within the same organization.

But more importantly – educational content representations are not uniform documents. These documents usually consist of completely different and semi-independent parts. For example, generic course programme may include a generic text description, a multilevel list, describing the topics within the course (each of which may or may not come with a short text description of its own) and table of learning outcomes (which themselves consist of an action verb from a small set of words and a learning outcome made using various terms bounded only by course's field of knowledge).

It is apparent, that different parts of the document require individual approach and different algorithms could show better results for these parts. For instance analysis of aforementioned learning outcomes would involve different approaches for its action verb and learning statement parts. While learning statement analysis falls under short text semantic similarity calculation and has to be solved by NLP methods, same cannot be said for the action verbs. Action verbs come from a small taxonomical list. One of the researchers [55], suggested a relatively simple algorithm for semantic similarity computation between these – by using a specific matrix representing position of the verbs in a space defined by their cognitive process domain and complexity coordinates. The similarity in this approach is determined by the distance between specified verbs in this matrix.

Ontology construction is one of the important tasks that needs to be solved when it comes to any problem involving knowledge extraction. Ontologies allow us to set internal structure to the knowledge represented by more or less plain text. For educational content, specifically course programmes and educational programmes, there exists the need to structure both the internal relations between smaller didactical units, and the way those units cover the underlying knowledge. By doing so, we can compare programmes not just as documents describing them, but as actual objects with their explicit internal structures. An example of research in this area, would be the curriculum and syllabus ontologies, suggested in [56]. Along with those, researchers also described a general algorithm for mapping syllabus to the specific knowledge units and classifying it through the usage of the ontology of underlying knowledge. Another researcher [57] suggested a general algorithm for ontology construction based on three-phased didactical activity model: teaching, learning and examination. This algorithm, in fact, produces multiple ontologies, such as Course Basic Subject Ontology, that includes basic notions and concepts of the course, Course Practical Activities Ontology, which covers a variety of terms related to practical activities for the corresponding chapters of the course and Basic Examination Ontology, which specifies terms related to the student evaluation process.

Another example of extracting concepts from educational documents according to some sort of a structure is represented in research [58]. Authors present approaches to classifying examination questions into the concept hierarchy for underlying knowledge to determine what exactly the question evaluates. The proposed model draws inspiration from widely known hierarchical classification techniques.

In [59], authors described a concept of an intelligent system for analysis and synthesis of educational content which took labor market into account. That system's architecture in context of our recent research is shown in Figure 3. This novel approach is intended to solve the complex problem of decreasing the amount of manual labor that goes into preparing educational content by Russian educational organizations in particular, while at the same time improving the overall quality of the content itself.

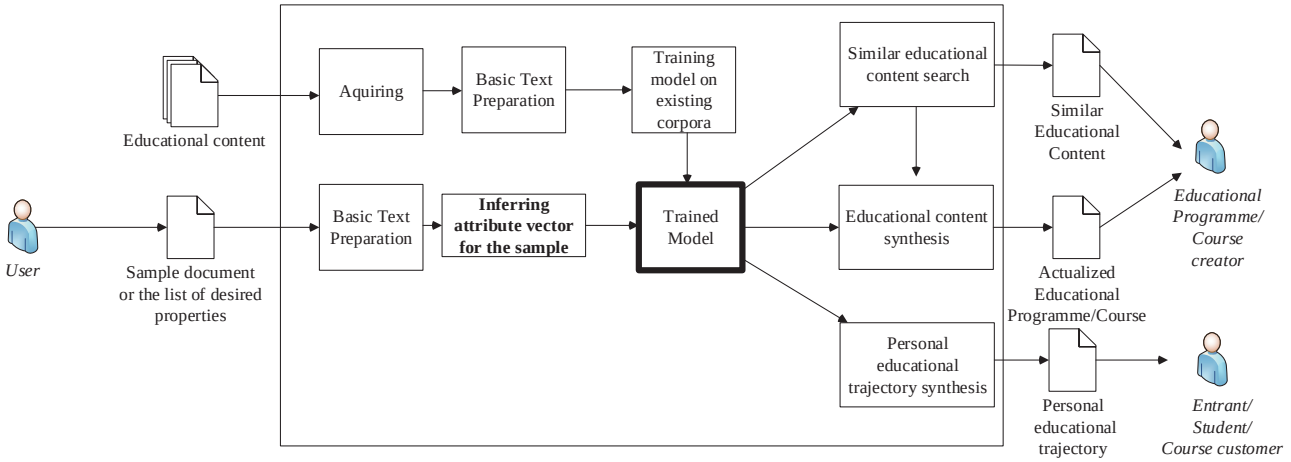


Figure 3 – General architecture of the proposed educational content synthesis system

The general workflow of the system is focused on acquiring educational content to put together a corpora and train the system on it, as to be able to embed the documents into the vector space. After it is done, the system would be used to infer the vectors for new documents coming in and compare those with the already known vectors from the corpora. This way we would achieve the possibility of search and synthesis of educational content, based on already known sources and new incoming samples. Here, by samples we understand both actual documents and more or less specific (potentially – also generated with the aid of the system) description of the desired content to be found or synthesized.

In order to increase effectiveness of new content generation and older content actualization (meaning educational programme, course programme and individual learning trajectories) it was suggested to solve the following tasks of educational content analysis:

- information retrieval of educational content, filtered by a set of criteria (learning outcomes, labor market requirements, requirements to content contents, educational and professional standard requirements) sorted according to their relevance;
- calculate structural and semantic similarity of educational content;
- automatic generation of knowledge base;
- variative structure and contents generation from existing knowledge base according to specified criteria;

To resolve the task of educational content analysis using the machine learning approaches one needs to:

- study more common formats of educational content representation;
- choose a set of criteria for each type of educational content;
- create a document corpora of educational content and prepare input data for further analysis;
- choose machine learning models (possibly hybrid) and algorithms and develop them, utilizing tools and libraries;
- train chosen models on created corpora and evaluate their effectiveness;
- choose the most effective algorithm (or combination) for each task of educational content analysis.

## Conclusion

In this paper we reviewed modern approaches to solution of a variety of NLP tasks. Nowadays, the most effective approaches to text analysis in both English is Russian tend to be based on distributional semantics, utilizing neural networks combined with semantic-syntactic analysis, using var-

ious deep learning architectures, such as recurrent neural networks and convolutional neural networks. We propose to evaluate effectiveness of those methods in regard to educational content analysis and synthesis in a future research.

## References


[1] ***Damashek M.*** Gauging similarity with n-grams: Language-independent categorization of text // Science, New Series, 1995, vol. 267, pp. 843-848.

[2] ***Harris Zellig S.*** Distributional Structure [Journal] // WORD, 1954, no. 10, pp. 146-162.

[3] ***Jones K.*** A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation, 1972, no. 28, pp. 11-21.

[4] ***Manning C., Raghavan P., Schutze H.*** Scoring, term weighting, and the vector space model. Introduction to Information Retrieval, 2008, 100 p.

[5] ***Panchenko A., Morozova O., Naets H.*** A Semantic Similarity Measure Based on Lexico-Syntactic Patterns. Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), Vienna (Austria), 2012, pp. 174-178.

[6] ***Osipov GS., Smirnov IV., Tihomirov IA., Shelmanov A.*** Relational-situational method for intelligent search and analysis of scientific publications // Proceedings of the Integrating IR Technologies for Professional Search Workshop, 2013, pp. 57-64.

[7] ***Panchenko A.*** Similarity Measures for Semantic Relation Extraction. PhD thesis. Universit´e catholique de Louvain, 2013, p. 197.

[8] ***Sobhana N., Pabitra M., Ghosh SK***. Conditional Random Field Based Named Entity Recognition in Geological text. Journal of Computer Applications, 2010, no. 2, vol. 1 (3), pp. 119-125.

[9] ***McCallum A., Li Wei.*** Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the 7th Conference on Natural Language Learning, 2003, pp. 188-191.

[10] ***Settles B.*** Biomedical named entity recognition using conditional random fields and rich feature sets. // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications, 2004, pp. 104-107.

[11] ***Galley M.*** A skip-chain conditional random field for ranking meeting utterances by importance // Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2006, pp. 364-372.

[12] ***Shen D., Sun J.-T.,Li H., Yang Q., Chen Z.*** Document summarization using conditional random fields. Proceedings of the 20th international joint conference on Artifical intelligence, 2007, pp. 2862-2867.

[13] ***Nothman J.*** Learning multilingual named entity recognition from Wikipedia. // Journal Artificial Intelligence archive, 2013, vol. 194, pp. 151-175.

[14] ***Yao X., Durme B.*** Information extraction over structured data: Question answering with freebase // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, vol. 1, pp. 956-966.

[15] ***Irmak U., Kraft R.*** A scalable machine-learning approach for semi-structured named entity recognition // Proceedings of the 19th international conference on World wide web, 2010, pp. 461-470.

[16] ***Jiang M.*** A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries // J Am Med Inform Assoc., 2011, no. 18(5), pp. 601–606.

[17] ***Ritter A., Clark S., Mausam, Etzioni O.*** Named entity recognition in tweets: an experimental study // Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1524-1534.

[18] ***Sidorov G., Velasquez F., Stamatatos E., Gelbukh A., Chanona-Hernandez L.*** Syntactic N-grams as Machine Learning Features for Natural Language Processing. 11th Mexican International Conference on Artificial Intelligence (MICAI), 2012, vol. 7630, pp. 1–11.

[19] ***Jursic M., Mozetic I., Erjavec T., Lavrac N.*** LemmaGen: Multilin-gual Lemmatisation with Induced Ripple-Down Rules // Journal of Universal Computer Science, 2010, no. 16, pp. 1190–1214.

[20] ***Liu H., Hunter L., Keselj V., Verspoor K***. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. // PLoS ONE, 2013, no. 8 (4).

[21] ***Percha B., Garten Y., Altman RB.*** Discovert and Explanation of drug-drug Interactions via text mining. // Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing, 2012, pp. 410-421.

[22] ***Palomino-Garibay A., Camacho-Gonzalez AT., Fierro-Villaneda RA., Hernandez-Farias I., Buscaldi D., Meza-Ruiz IV.*** A Random Forest Approach for Authorship Profiling. // CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, Toulouse, France, 2015.

[23] ***Wu Q.; Ye Y., Zhang H., Ng M. K., Ho S.-S.*** ForesTexter: An efficient random forest algorithm for imbalanced text categorization. // Journal Knowledge-Based Systems archive, 2014, vol. 67, pp. 105-116.

[24] ***Stoyanov V., Cardie C., Gilbert N., Riloff E., Buttler D., & Hysom D.*** Coreference Resolution with Reconcile. *Proceedings of the ACL 2010 Conference Short Papers*, (pp. 156-161).
[25] ***Fader, A., Zettlemoyer, L., & Etzioni, O.*** Paraphrase-driven learning for open question answering. *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, (pp. 1608-1618). Sofia, Bulgaria.
[26] ***Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J.*** Distributed Representations of Words and Phrases and their Compositionality. // Advances in neural information processing systems, 2013, pp. 3111-3119.
[27] ***Bengio Y., Ducharme R., Vincent P., Janvin C.*** A neural probabilistic language model. // The Journal of Machine Learning Research, 2003, no. 3, pp. 1137–1155.
[28] ***Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean***. Efficient estimation of word representations in vector space. ICLR Workshop, 2013.
[29] ***Pennington Jeffrey, Socher Richard u Manning Christopher D.*** GloVe: Global Vectors for Word Representation, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp. 1532–1543.
[30] ***Panchenko A., Loukachevitch NV., Ustalov D., Paperno D., Meyer CM., Konstantinova N.*** RUSSE: The First Workshop on Russian Semantic Similarity [Report]. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue", Moscow, RGGU, 2015, vol. 2, pp. 89-105.
[31] ***Arefyev NV., Panchenko AI., Lukanin AV., Lesota OO., Romanov PV.*** Evaluating Three Corpus‑based Semantic Similarity Systems for Russian. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue", Moscow, RGGU, 2015, vol. 2, pp. 106-118.
[32] ***Quec Le, Tomas Mikolov.*** Distributed Represenations of Sentences and Documents. In Proceedings of ICML 2014, pp.1188–1196.
[33] ***Bengio Yoshua.*** Learning Deep Architectures for AI // Journal Foundations and Trends in Machine Learning, 2009, no.1, vol. 2, pp. 1-127.
[34] ***Wang, Hao.*** Semantic Deep Learning, University of Oregon. 2015. - 42 p.
[35] ***Mesnil G., He X., Deng L., Bengio Y.*** Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. // 14th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2013, pp. 3771-3775.
[36] ***Socher Richard***. Recursive Deep Learning for Natural Language Processing and Computer Vision: Ph.D. thesis. Stanford University, 2014, 204 p.
[37] ***Cho K., Van Merrienboer B., Gulçehre Ç., Bahdanau D., Bougares F., Schwenk H., Bengio Y.*** Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724-1734.
[38] ***Bahdanau D., Cho K., Bengio Y***. Neural machine translation by jointly learning to align and translate. // Proceedings International Conference on Learning Representations, 2015. - 15 p.
[39] ***Skorinkin D.A., Budnikov E., Stepanova M., Matavina P., Chelombeeva A***. Information Extraction Based on Deep Syntactic-Semantic Analysis // Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue", Moscow, RGGU, 2016, pp. 721-733.
[40] ***Bogdanov AV., Dzhumaev SS., Skorinkin DA., Starostin AS.*** Anaphora Analysis based on ABBYY Compreno Linguistic Technologies. // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue 2014", Moscow, RGGU, 2014, pp. 89-101.
[41] ***Tymoshenko K., Moschitti A.*** Assessing the impact of syntactic and semantic structures for answer passages reranking. // Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015, pp. 1451-1460.
[42] ***Yih Wen-tau S., He X., Meek C***. Semantic Parsing for Single-Relation Question Answering. // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, Maryland, USA, 2014, pp. 643–648.
[43] ***Kalchbrenner N., Grefenstette Ed., Blunsom P.*** A Convolutional Neural Network for Modelling Sentences // arXiv preprint arXiv:1404.2188, 2014.
[44] ***Dos Santos Cicero, Gatti Maira***. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. // Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014), 2014.
[45] ***Irsoy O., Cardie C.*** Opinion Mining with Deep Recurrent Neural Networks. // Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014, pp. 720-728.
[46] ***Loukachevitch NV., Rubtsova YV.*** SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis. // Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue", Moscow, RGGU, 2016, pp. 416-427.
[47] ***Arkhipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D.*** Comparison of Neural Network Architectures for Sentiment Analysis of Russian Tweets. // Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue", Moscow, RGGU, 2016, pp. 50-58.

[48] ***Prata D., Baker R., Costa E., Rose C., Cui Y.*** Detecting and understanding the impact of cognitive and interpersonal conflict in computer supported collaborative learning environments.// International Conference on Educational Data Mining, 2009, pp. 131-140.
[49] ***Song D., Lin H., Yang Z.*** Opinion Mining in e-Learning System // International Conference on Network and Parallel Computing Workshops, 2007, pp. 788-792.
[50] ***Zhang X., Mostow J., Duke N., Trotochaud C., Valeri J., Corbett A.*** Mining Free-form Spoken Responses to Tutor Prompts. // International conference on Educational Data Mining, 2008, pp. 234-241.
[51] ***Chiu DY., Pan YC., Chang WC.*** Using rough set theory to construct e-learning faq retrieval infrastructure. // IEEE Ubi-Media Computing Conference, 2008, pp. 547-552.
[52] ***Saini P. S., Sona D., Veeramachaneni S., Ronchetti M.*** Making E-Learning Better Through Machine Learning // International Conference on Methods and Technologies for Learning, 2005, pp. 1-6.
[53] ***McNamara DS., Crossley SA., Roscoe RD.*** Natural language processing in an intelligent writing strategy tutoring system. Behavior Research Methods, 45, 2013, pp. 499--515.
[54] ***Sung et al.*** Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning // Behavior Research Methods, 47 (2) (2015), pp. 340–354.
[55] ***Chernikova E***. A Novel Process Model-driven Approach to Comparing Educational Courses using Ontology Alignment, 2014. - http://hdl.handle.net/2086/10107.
[56] ***Chung H., Kim J.*** An Ontological Approach for Semantic Modeling of Curriculum and Syllabus in Higher Education // International Journal of Information and Education Technology vol. 6, no. 5, pp. 365-369, 2016.
[57] ***Oprea M.*** On the Use of Educational Ontologies as Support Tools for Didactical Activities, Proceedings of the International Conference on Virtual Learning(ICVL2012), Nov. 2012, pp. 67-73.
[58] ***Foley J., Allan J.*** Retrieving Hierarchical Syllabus Items for Exam Question Analysis // Advances in Information Retrieval, March 2016, pp. 575-586.
[59] ***Botov D., Klenin J.*** Educational Content Semantic Modelling for Mining of Training Courses According to the Requirements of the Labor Market. // In Proceedings of the 1st International Workshop on Technologies of Digital Signal Processing and Storing, Russia, Ufa, UGATU, 2015, pp. 214-218.

---

# ОБ ИСПОЛЬЗОВАНИИ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА НА ПРИМЕРЕ АНАЛИЗА ОБРАЗОВАТЕЛЬНОГО КОНТЕНТА

**А.В. Мельников[1], Д.С. Ботов[2], Ю.Д. Кленин[3]**

*Челябинский государственный университет, Институт информационных технологий, Челябинск, Россия*
*mav@csu.ru[1], dmbotov@gmail.com[2], jklen@yandex.ru[3]*

**Аннотация**

Рассмотрены наиболее популярные подходы к различным задачам обработки естественного языка (NLP), преимущественно использующие машинное обучение: от классических до передовых технологий. Большую часть подходов можно разделить на три подмножества. В одном - используют гипотезу дистрибутивной семантики, в другом - информацию из графовых баз знаний (например, онтологий), и в третьем - анализируют лексико-синтаксические шаблоны в документах. Основной фокус статьи на первом из этих подходов. Один из наиболее важных подготовительных шагов NLP – это задача представления документов в виде числовых векторов. Существуют различные методы, начиная от простейшей модели “Мешок Слов” и заканчивая изощрёнными подходами к машинному обучению, например вложению слов. На сегодняшний день в задаче поиска информации самое высокое качество и для английского, и для русского языков достижимо подходами на основе алгоритмов вложения слов, тренированных на тщательном подборе корпусов в сочетании с синтаксическим и семантическим анализом на основе различных глубоких нейронных сетей. Различные алгоритмы машинного обучения используются в задачах NLP таких как тегирование частей речи, реферирование текстов, распознавание именованных сущностей, классификация документов, извлечение тем и отношений сущностей, и вопросно-ответные системы на естественном языке. Рассмотрена применимость данных алгоритмов к анализу образовательного контента, а также предложен подход к приложению возможностей NLP и машинного обучения к анализу и синтезу образовательного контента в виде системы поддержки принятия решений.

***Ключевые слова:*** *машинное обучение, обработка естественного языка, анализ образовательного контента, семантическая близость, глубокое обучение, нейронные сети.*

***Цитирование:*** *Мельников А.В.* Об использовании машинного обучения в задачах обработки естественного языка на примере анализа образовательного контента [In English] / А.В. Мельников, Д.С. Ботов, Ю.Д. Кленин // Онтология проектирования. – 2017. – Т. 7, №1(23). - С. 34-47. – DOI: 10.18287/2223-9537-2017-7-1-34-47.

## Сведения об авторах

***Мельников Андрей Витальевич***, родился 22 января 1956 года, окончил Челябинский политехнический институт по специальности "Автоматика и телемеханика", доктор технических наук. Профессор кафедры информационных технологий и экономической информатики Института информационных технологий Челябинского государственного университета. Преподаваемые дисциплины: анализ информационных технологий, информационные технологии и системы в бизнесе, методология научных исследований, сети и телекоммуникации. Автор 7 статей по тематике интеллектуального анализа образовательного контента.

***Melnikov Andrey Vitalievich***, born 22nd of January 1956, specialization: electrical engineer ("Automation and remote control ", Chelyabinsk State University), Doctor of Science. Professor of IT and Economical Computer Science Department of Informational Technologies Institute at Chelyabinsk State University. Courses: IT analysis, Informational Technologies and Systems in Business, Scientific Research Methodology, Intro to Specialization, Networks and Telecommunications. He is author of 7 articles about intelligent analysis of educational content.

***Ботов Дмитрий Сергеевич***, родился 23 июня 1989 года, окончил Южно-Уральский государственный университет по специальности "Вычислительные машины, комплексы, системы и сети". Старший преподаватель кафедры информационных технологий и экономической информатики Института информационных технологий Челябинского государственного университета. Преподаваемые дисциплины: базы данных, объектно-ориентированный анализ и программирование, программирование на языке Java, управление жизненным циклом информационных систем, программная инженерия. Автор 5 статей по тематике интеллектуального анализа образовательного контента.

***Botov Dmitry Sergeevich***, born 23rd of June 1989, specialization: engineer of "Computers, Complexes, Systems and Networks" (South Ural State University). Senior Lecturer of of IT and Economical Computer Science Department of Informational Technologies Institute at Chelyabinsk State University. Courses: Databases, Object-Oriented Analysis and Programming, Java Programming, Informational System Life Cycle Management, Software Engineering. He is author of 5 articles about intelligent analysis of educational content.

***Кленин Юлий Дмитриевич***, родился 1 февраля 1994 года, окончил Южно-Уральский государственный университет по специальности "Информатика и вычислительная техника". Проходит обучение в магистратуре по специальности "Фундаментальная информатика" (Институт информационных технологий Челябинского государственного университета). Автор 3 статей по тематике интеллектуального анализа образовательного контента.

***Klenin Julius Dmitrievich***, born 1st of February 1994, specialization: bachelor of "Computer Science and Computing" (South Ural State University). Studies for master's degree in "Fundamental Computer Science" (Informational Technologies Institute at Chelyabinsk State University). He is author of 3 articles about intelligent analysis of educational content.