Processing of Natural Language Texts in Information Learning Systems: Ontological Approach*

Olha Tkachenko[1,2,*,†], Kostiantyn Tkachenko[3,†] and Oleksandr Tkachenko[3,†]

[1] State University of Infrastructure and Technologies, 9 Kirillivska str., 04071 Kyiv, Ukraine
[2] Borys Grinchenko Kyiv Metropolitan University, 18/2 Bulvarno-Kudryavska str., 04053 Kyiv, Ukraine
[3] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute," 37 Beresteyskyi ave., 03056 Kyiv, Ukraine

Abstract

Effective organization of learning processes supported by relevant information learning systems consists of choosing the appropriate technology for analyzing natural language text of educational content that can ensure: individualization of learning; adequately adapt educational content to students; support the so-called "understanding" of texts in Ukrainian and English (these texts are provided to students as fragments of the educational content of the course supported by the relevant information learning system, as well as by the students themselves in the learning process (description of the solved problem; answers given in their own words rather than selected from answer options; tests, questions to the system, etc.)); creation of prototypes; continuous iteration in recognition and processing of natural language texts; maximum reliability and efficiency of learning processes. The article considers an ontological approach to a formalized description of knowledge in various subject areas of information learning systems and the essence of linguistic analysis of texts of educational content provided in a natural language (Ukrainian and/or English). The article analyzes modern methods of organizing learning processes based on the perception ("understanding") of information (provided by natural language texts) by students. The results of the analysis were used in the development of a software product to support the educational/training/educational process in Ukrainian/English, which improves the efficiency of learning processes based on the technology of natural language processing of educational content. The paper shows modern methods of linguistic analysis of natural language texts. The analysis of tokenization, normalization, stemming, and lematization methods is carried out. Their use in information learning systems in the linguistic analysis of many natural language texts (fragments) of educational content is considered.

Keywords

natural language text processing, educational content, information learning system, ontology

1. Introduction

Modern information technologies are integrated into various spheres of life of both individuals and society as a whole. One of such spheres is education. That is why the integration of modern information technologies for working with text information (educational content of online courses, students' answers provided in natural language, etc.) determines:

- The ontological approach to organizing learning processes.
- Increasing the level of individualization of educational content provided to students.
- Understanding the system of students' answers/questions provided by them when communicating with the system.

Natural Language Processing (NLP) technologies [1, 2]. Thanks to the use of artificial intelligence, neural networks [3–5], machine learning, and ontological modeling [5–8] are increasingly penetrating learning processes, expanding their capabilities due to:

*CPITS 2025: Workshop on Cybersecurity Providing in Information and Telecommunication Systems, February 28, 2025, Kyiv, Ukraine
* Corresponding author.
† These authors contributed equally.
oitkachen@gmail.com (O. Tkachenko); tkachenko.kostyantyn@gmail.com (K. Tkachenko); aatokg@gmail.com (O. Tkachenko)
0000-0003-1800-618X (O. Tkachenko); 0000-0003-0549-3396 (K. Tkachenko); 0000-0001-6911-2770 (O. Tkachenko)

- Individualization of learning.
- Improving interaction between online course users: students, teachers, authors of educational content, methodologists, representatives of higher education management, faculty administration, etc., and the corresponding information learning system.

It should be noted that information learning systems that use NLP technologies face problems associated with incomplete and/or false data, which may lead, in particular, to:

- Building models of educational content, for example, semantic and neural networks and hierarchical and ontological models the use of which, with an inadequate assessment of the level of initial competencies of students, leads to the formation of ineffective (and sometimes completely unnecessary for the student) learning trajectories.
- Building models of student communication with the system, the use of which, with an incorrect understanding of the answers/requests from students, leads to providing the wrong fragments of individualized educational content (or its fragments) and incorrect assessment of student answers/questions.

Modern information learning systems should use many technological solutions for natural language processing. That is why there is a gradual transition to learning that supports the so-called "free" expression of opinion by students who describe their version of solving a problem or answering questions using natural language texts. Processing of natural language text in information learning began when A. Turing proposed a model for testing the system for so-called "consciousness." Such methods as normalization fragmentation and tokenization of text perform pre-processing of text.

The UIMA (*Unstructured Information Management Architecture*) platform [9] plays an important role in the semantic analysis of natural language texts. This platform, in particular, is used to build systems for semantic analysis of unstructured information, unifies the process of processing natural language texts, and allows for analyzing multimedia files. Applications created on the UIMA platform have a multi-component architecture, where each component performs specific functions [9]: language identification, syntactic analysis, and direct annotation of the text.

NLP is already widely used in chatbots, but it should be noted that this mainly applies to texts in English. The development of similar NLP methods and algorithms to texts in the Ukrainian language is relevant and requires its solution.

2. Classification of natural language texts

One of the current tasks in the process of processing (in particular, analysis) of natural language texts is their preliminary classification, which involves assigning the text to one or more thematic sections. Most methods are designed to search for documents on the Internet by keywords. When using Big Data technology, a large array of documents is pre-processed and loaded into a database and knowledge base. In the texts, it is necessary to determine the keywords by which they can be assigned to one or more thematic sections. The development of models and methods for classifying natural language texts is a current problem in our time.

The purpose of the work is to study methods for analyzing, processing, and classifying natural language texts and developing a corresponding information system. The problem of pattern recognition and automated analysis, processing, and classification of natural language texts is considered in [10]. In this work, various types of classifiers are investigated, in particular. The classic classification method is the TF-IDF method [11], which is based on the so-called "vector text model" and is currently the most effective, widespread, and used in information retrieval systems.

There are many methods for analyzing, processing, and classifying natural language texts, which use, in particular, neural networks, clustering methods, and porn vectors. Classification of natural language texts for predicting the thematic category of natural language text in the English-language

Wikipedia based on the use of the Apache Spark platform, which is built into the Big Data Hadoop system, is considered in [12]. Various aspects of working with Big Data are considered in [13]. To obtain the necessary knowledge, such methods of data classification and clustering as *k*-Means, Support Vector Machine, Naive Bayes, k-Nearest Neighbor, Map Reduce, and Apache Spark are used.

The classification system works as follows: in a set of natural language texts, each text is a member of the i^{th} collection of texts

$$T_i \subset T \; (T = ¿ i=1 ¿ n T_i)$$

and is a member of the possible categories

$$C_j \subset C \; (C = ¿ j=1 ¿ m C_j).$$

Then the classification is the operation of matching each text with one or more of their classes.

The paper considers a classification that uses the following categories: education, entertainment, culture, history, and world. The classification system attributes the text to a certain category (class) C_j. To correctly attribute the text to a certain class, the linguistic text analysis system must have, in particular, such information as the keywords included in the text. When making decisions about assigning a text to a certain class, the system does so based on the vector of information features, which are the occurrences of a certain set of keywords (terms) in the text:

$$KW_{T_i} = \{KW_{1T_i}, KW_{2T_i}, ..., KW_{\mathrm{IT}_i}\}.$$

The feature vector of the text T_i is

$$X_i = \{x_{i1}, x_{i2}, ..., x_{il}\},$$

where x_{ik} takes the value 0 if the term x_{ik} is not included in the text and 1 if it is.

The classification system (based on the results of the classification of the previous T_{i-1} texts) attributes the text T_i to one (sometimes to several) of the text classes. The algorithm for classifying natural language texts has the form: of lexical analysis (parsing) of the text we read the text, divide the text into pairs <keyword—meaning>, and count the occurrences of keywords (terms, concept) in the text. After lexical analysis of the text, significant words are searched for. To assess significance, methods are used that take into account the frequency of occurrence of terms and characteristics that reflect the common frequency of occurrence (association) and the density of distribution of terms in the text.

We count the number of occurrences of each word in the text. We calculate the frequency rank of the word according to the formula

$$Cn = P \times R$$

where P is the probability of detecting a word, R is the frequency rank of the word, Cn is const.

To do this:

- We find the most popular words by counting words to create tuples of the type (word, counter).
- We find significant terms.
- We form a sequence of significant words (by their rank).
- Based on the found significant words and their ranks, the system decides to assign the text to the restored class Cj.
- The main point of the considered approach to classifying natural language texts is to manage the set of texts.

3. Natural language text processing system

Nowadays, computational linguistics has become one of the most important areas of artificial intelligence, the significant results of which are lexicographic systems, electronic dictionaries, machine translation systems, and automatic abstracting systems. At the same time, the problem of intelligent natural language text processing systems is the difficulty of establishing the correct mapping of the actual semantic-syntactic structure of a sentence into its internal logical representation, which is automatically generated by the system.

Most modern models used in natural language text processing are "isolated" in structure. Examples of natural language text semantic models are semantic networks and frames. Different levels of natural language text processing are combined algorithmically (functionally). The syntax of natural language sentences is expressed using:

- Chomsky grammars
- Halliday grammar
- Extended transition network
- Lexical-morphological models (these models are structurally isolated at their levels).

All levels of natural language are interconnected not only functionally, but also structurally. The system for processing natural language texts is based on the principles of structuring data presented in the corresponding knowledge base. The knowledge base uses an ontological-semantic network that describes objects, properties, and relations of the corresponding subject area [6]. Of great importance in the linguistic analysis of natural language texts is the tree of actions and relations, which describes the hierarchy of actions—from abstract actions to specific representatives-subclasses, which allows using the inheritance mechanism when describing actions and their properties.

In ontology, concepts of the action type are considered functions with a set of arguments that correspond to various aspects of the action [5]. Each argument is specified by the subject area of the definition. For each aspect of the action, it is possible to:

- Specify the semantic network of the input natural language text (e.g., a fragment of educational content).
- Specify the type of concept used if knowledge is required from the knowledge base of the corresponding problem subject area.
- Specify the syntactic structure of the natural language text that should be used.

This way, it is possible to obtain a structural connection between the semantics and syntax of the natural language text. In ontology, the objects, properties, and relations of the problem subject area are described semantically. Therefore, lexical units (which correspond to these objects, properties, and relations) can be stored directly in the ontology using special nodes used as names of objects, properties, and relations in a certain natural language (e.g., the natural language in which the educational content of a certain course/topic is presented, etc.). The resulting structure will link the semantics, syntax, and vocabulary of the natural language. This simplifies the procedures for synthesizing and analyzing natural language texts and increases their efficiency. The main functions of the automatic/automated natural language text system are analysis (translation of natural language text into a formal logical representation of its meaning) and synthesis (generation of natural language text based on the logical representation of information). Automaticity or automation depends on the language of the problem subject area. An important part of the core of the natural language text processing system is the knowledge base based on semantic ontology. It is used at all stages of analysis, and the results of natural language text analysis are presented as a semantic network with the ability to add consistent facts obtained from the input text to the knowledge base. Analysis stages are morphological and lexical analysis, syntactic analysis, and semantic analysis.

Automatic indexing and abstracting involve processing the facts obtained as a result of analysis from the context knowledge base. The text synthesis subsystem, using rules from the ontology and templates of the syntactic analysis subsystem, builds linear sentence structures based on the semantic network of the natural language text. These structures are then filled with the corresponding lexemes in the required form using the lexical analysis subsystem. Semantic ontology is a directed hypergraph, each node of which represents a concept and has a set of links-relationships between this node-concept and other nodes-concepts.

Each node has a name—a word that characterizes the meaning of the node. The most important type of link in the graph is the "to be" relationship. Links-relationships form an ontological hierarchical graph (tree) of natural language concepts, the root of which is the most abstract object "all". The nodes "action", "object", "property" and "relationship" are natural language categories distinguished in the ontology, having several less abstract objects ("sons"). Each node can have several fathers, that is, it can inherit semantic relations and attributes of all its fathers [7, 14].

The knowledge base is an ontological hierarchical network [8] that contains a set of language concepts. In addition to the vertical "to be" links, it contains a set of horizontal relationships ("has the property", "does", etc.) that describe their objects, relations, actions, and other facts of the problem area. Hierarchy ensures the efficient use of the inheritance mechanism, which helps to avoid redundancy. When adding a new problem area to the system, its hierarchical network is added, which describes its concepts and the relationships between them. This explains the relative ease of adding new topics to the system. The lexicon is linked to the semantic ontology—a specific word to the corresponding concept. In the case of synonymy, one concept corresponds to several words of the lexicon, in the case of homonymy—one word corresponds to several concepts.

For morphological and lexical analysis of natural language text is presented as a sequence of natural language sentences. The task of the morphological and lexical analyzer is to find an entry in the lexicon of the system ontology for each word of the input sentences and fully determine the morpholexical characteristics of input words (gender, number, case, etc.). To solve this problem, word-forming models were developed for the English and Ukrainian languages. Syntactic analysis of natural language texts: linear sequences of natural language sentences explicitly contain all morpholexical characteristics.

These sequences are transformed into syntactic structures based on syntactic templates based on the corresponding ontology. During the first pass of the syntactic analyzer over the natural language text, syntactic groups (verb groups, noun groups, etc.) are formed. During the second pass, individual groups are assembled into a single syntactic structure. This assembly is performed by filling in the aspect fields of the verb group according to the syntactic templates attached to the ontology. Sometimes the syntactic analyzer cannot unambiguously determine the correct syntactic structure of a natural language sentence using syntactic rules. For example, the sentence "The player moves the figure in the game on the field". The rules of syntax cannot answer what the lexeme "field" is associated with the lexeme "game" or the lexeme "player". This is a question of semantics. The semantic analyzer calculates the length of paths in the semantic ontology between one pair of concepts and between another (from "player" to "field" and from "game" to "field"). After comparing the lengths of the paths, the appropriate conclusion is made. Then the natural language text is returned to the syntactic analyzer, which completes the formation of the syntactic structure of the NL sentence.

Semantic analysis of natural language texts works in parallel with syntactic analysis. The semantic analyzer replaces words with concepts in the collected structures. It adds a semantic context from the semantic network along with the concept—a set of specific attributes and relationships of the concept. Then the concepts combined in the structure are checked for consistency.

After this, a logical check is performed on how naturally the objects and relations that connect them are combined in the structure and how the formed network of the NL sentence is isomorphic to the semantic ontology of the subject area implemented in the system. First, pronouns are replaced with the concepts to which they refer in the natural language text.

The corresponding algorithm is guided by morpholexical characteristics of pronouns (gender, number, case) and word concepts that were encountered in the text (they must match) and semantic properties—the position of the pronoun in the semantic network of the sentence, which must be similar to that occupied by the concept candidate for its place in the semantic network of the system. Thus, a semantic network of the entire natural language text can be obtained from the networks of sentences. Generation of the abstract of the natural language text: the nodes of the semantic network of the natural language text can be "weighed." The most important nodes of the network are considered to be the vertices that have the greatest number of connections with others. By weighing the vertices of the graph and discarding the lightest ones, we obtain a semantic image of the future abstract for the corresponding natural language text.

Having carried out a comparative analysis of the concepts and connections of the obtained image with the networks of the subject area (which are contained in the semantic ontology), the abstract generator concludes the topic of the natural language text and determines the category to which the text belongs.

In the resulting optimized graph, the vertices and connections have their assessment. In the simplest case, it corresponds to the order in which the sentences corresponding to these concepts appear in the text. The text generator sequentially processes the subgraphs of the network, the vertices and connections of which have the same time assessment in ascending order—from the smallest to the largest. Using the syntactic analyzer, the abstract generator finds a correspondence between the structure of the graph/subgraph and a certain syntactic template. By the found syntactic patterns, the generator reconstructs the graph/subgraph structure into a linear one. Then, using the morphological-lexical analyzer, the lexemes corresponding to the concepts are inserted in the required form into the positions of the resulting linear structure. All subgraphs of the optimized semantic network of the natural language text are processed. After that, a text summary is generated. To improve the quality of the text, it is advisable to use mechanisms of synonyms, pronouns, and other stylistic devices.

4. Processing of natural language texts using neural networks

The process of determining the content of a natural language text test can be automated using AI. The Bayesian method [15] or SVM [13] was initially used as the basic method. The development of neural networks convolutional neural networks (CNN) and recurrent neural networks (RNN) led to their use in computational linguistics. First of all, determining the content of natural language texts concerns the process of Natural Language Processing [1, 2].

Let us consider some steps to solving NLP problems: TF-IDF characteristic. To process natural language text in the system, it must acquire a quantitative form. There are several methods for such conversion: TF-IDF [16, 17]. TF (*term frequency*)—the frequency of each used word (term, concept). IDF (*inverse document frequency*)—the inverse number of terms in the natural language text. The TF-IDF indicator indicates how rare a certain term is. For example, interjections, conjunctions, or exclamations will be the most common, and, accordingly, will have a low TF-IDF.

TF-IDF characteristics allow you to rank terms. For this, Word Embedding technology is used, which maps words or phrases into vectors of real numbers.

The formal technology mentioned is a set of various methods, in particular, GloVe (*Global Vectors* —an algorithm for converting unlabeled data (terms) into continuous vectors [18].

GloVe vectors (pre-trained on data from Wikipedia and Gigaword 5) capture the semantics of sentences well.

But this algorithm is aimed at texts in English, not Ukrainian. natural language text contains a lot of different information, in particular, the title, the text itself, images, the author, links to the source, etc. We will use only the title and text.

Computational linguistics. Natural language texts, individualized for a specific student, differ in the number of words and terms used. The range of words in detailed intermediate-level learning content is greater, but sentences are shorter than in high-level content. Given a set of m natural

language texts (fragments of learning content of different levels of complexity and detail), which can be represented as follows:

$$CL = \{CL_1^{L1}{}_{d1}, CL_1^{L1}{}_{d2}, \dots, CL_1^{L1}{}_{dk}, CL_2^{L1}{}_{d1}, CL_2^{L1}{}_{d2}, \dots, CL_2^{L1}{}_{dk}, \dots, CL_n^{L1}{}_{d1}, CL_n^{L1}{}_{d2}, \dots, CL_n^{L1}{}_{dk}, CL_1^{L2}{}_{d1},$$
$$CL_1^{L2}{}_{d2}, \dots, CL_1^{L2}{}_{dk}, CL_2^{L2}{}_{d1}, CL_2^{L2}{}_{d2}, \dots, CL_2^{L2}{}_{dk}, \dots, CL_n^{L2}{}_{d1}, CL_n^{L2}{}_{d2}, \dots, CL_n^{L2}{}_{dk}, \dots, CL_1^{Lp}{}_{d1}, CL_1^{Lp}{}_{d2}, \dots, CL_1^{Lp}{}_{dk},$$
$$CL_2^{Lp}{}_{d1}, CL_2^{Lp}{}_{d2}, \dots, CL_2^{Lp}{}_{dk}, \dots, CL_n^{Lp}{}_{d1}, CL_n^{Lp}{}_{d2}, \dots, CL_n^{Lp}{}_{dk}\} = CL_{jd_l}^{L_i}$$

where $j = 1, 2, \dots, n$; $i = 1, 2, \dots, p$; $l = 1, 2, \dots, k$, $m = n+k+p$, $CL_{jd_l}^{L_i}$ is j^{th} fragment of the learning content of the i^{th} level of complexity and l^{th} level of detail.

When determining the content of the natural language text, it is necessary to predict whether the student has mastered the corresponding fragment from *CL* or not.

In this case, the set of labels indicating the mastery of the corresponding fragment of educational content can be represented as follows:

$$\mu=\{1, 0\}^{\text{m}},$$

where 1 means that the mastery has occurred, and 0 means that it has not ($m = n+k+p$).

The set of functions $F_{jd_l}^{L_i}$ з

$$CL_{jd_l}^{L_i} \in CL$$

is obtained by syntactic analysis, in particular using TF-IDF and Word Embedding.

The accuracy of determining the individualized mastery of fragments of educational content using CNN and RNN is calculated using the following model for forming labels that will indicate the mastery of the corresponding fragment of educational content from *CL*:

$$\varphi: \{F_{jd_l}^{L_i}: T\}_{jl}^{i} \epsilon\, CL \rightarrow y,$$

where T is the set of all terms (concepts) of the natural language text (educational content or its fragment).

Let us describe the algorithm according to which data is prepared for linguistic analysis:

- "Text cleaning"—removing all non-letter expressions from the text (for example, numbers, commas, periods, and other punctuation marks) using a special library that provides access to regular expressions.
- Text analysis using methods that process natural language texts based on built-in word corpora:
  - Removing words from the set that do not carry an information load (for example, "and", "or", etc.); these words interfere with the correct analysis.
  - Removing linguistic variability, which is due to the use of morphemes, using the stemming operation (reducing a word to its base, for example, the words "learned" and "learning" will be replaced by the word "learn").
  - After this processing, the array of words is reduced to a set of bases.
  - Combination of stemming with lemmatization (reducing the word form to a lemma (normal dictionary form). For example, when processing the words "*bad*", "*worse*" and "*worst*" will have different bases, but the lemma of these words is the same—"*bad*".
- Checking the text for uninformative educational content after performing the above actions.
- Creating a dictionary necessary for the correct determination of the TF-IDF characteristic.
- Finding a frequency (frequency-polar) characteristic for each word.

Standard models are not always effective in classifying texts, so it is advisable to understand:

- Under CNN a neural network with a single-layer convolution and the appropriate setting of all its parameters.
- Under RNN an LSTM (long short-temp memory) [4, 19] network configured for text analysis, which has long-term and short-term memory.

The CNN model works faster than RNN but gives a less accurate classification result.

5. Learning with NLP-based recommendation systems

NLP-based recommendation systems significantly facilitate the learning process and increase the effectiveness of matching educational content, goals, methods and means of learning, and student profiles. This approach contributes to the automation and optimization of the learning process, ensuring accuracy, objectivity, and efficiency. The approach to recommendation systems based on the use of RNNs is described in [20].

Unlike traditional text vectorization methods, such as bag-of-words or TF-IDF, which do not take into account the context and word order, RNNs can more accurately detect semantic relationships in unstructured natural language texts. The RNN approach provides significantly higher accuracy in detecting semantic relevance between educational content and student knowledge level. The problem of processing natural language texts (for example, educational content of a discipline, answers, student requests, etc.) was studied in [1].

It should be noted that automatic analysis and comparison of natural language texts is often complicated by the language itself and the insufficient "development" of algorithms for processing such texts. Traditional methods, such as keywords or rules, often do not take into account the context and semantics, which leads to low accuracy. To solve this problem, an approach based on the BERT (*Bidirectional Encoder Representations from Transformers*) model [21] was proposed, which creates contextualized vector representations for words in sentences, taking into account their environment and dependencies. The possibilities of using methods for modeling educational content topics, such as Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA), are described in [1, 22] for the automatic classification of fragments by specialties, educational courses, types of tasks, and requirements for independent (individual) work, etc.

The LDA-based approach provided a more structured and informative representation of educational data. Applying LSA to the texts of educational content fragments, it is possible to automatically identify sets of competencies, skills, tools, methodologies, etc. that would be most relevant to specific requirements for a student to master the corresponding educational content. The use of LDA and LSA can be useful for more accurately identifying gaps in knowledge, competencies, skills, and abilities and building personalized recommendations for the student's professional development. Various NLP approaches and methods can be applied to improve recommendation systems in the learning process [5, 6, 17].

To increase the accuracy and efficiency of these systems, it is important to use advanced machine learning technologies, such as word embedding, neural networks (in particular, CNN, RNN), and transformers. These systems can analyze natural language texts, identify key skills and competencies, and provide personalized recommendations on the most relevant levels and details of educational content fragments for each student.

The main stages of recommendation generation, in particular, are:

1. Collection and preparation of data, which will be the basis for developing recommendations. The collected data contains noise, errors, and duplicates, so cleaning and filtering methods must be applied to eliminate these shortcomings.
   This can be done using deduplication, removal of incorrect records, and other data cleaning algorithms.
2. Extraction of key information from the requirements for mastering educational content, in particular, topic names, descriptions of competencies, skills and abilities, level of basic knowledge, etc.
3. Conducting pre-processing of the natural language text of educational content. This stage involves the use of various natural language processing techniques, for example, such as:

   - Tokenization (breaking down the text into individual words, terms, phrases, or symbols).
   - Text lemmatization (text normalization)—reducing words to the basic form.

Fig. 1 shows the result after tokenization and lemmatization of a fragment of educational content.

Text fragment
Artificial intelligence provides unique capabilities that seemed like fantasy a few years ago. Algorithms can recognize voices, identify people, detect errors, give recommendations, decode emotions, etc. Artificial intelligence successfully performs various monotonous operations and processes large volumes of data.
Tokenization and lemmatization
[‘Artificial’, ‘intelligence’, ‘unique’, ‘opportunities’, ‘fantasy’] [‘Algorithms’, ‘voice’, ‘people’, ‘errors’, ‘recommendations’, ‘decode’, ‘emotions’] [‘Artificial’, ‘intelligence’, ‘monotonous’, ‘operations’, ‘processes’, ‘large’, ‘volumes’, ‘data’,]

Figure 1: Result of tokenization and lemmatization of natural language text

- Removing unnecessary information (punctuation, punctuation marks, special characters, etc.).
- Cleaning the text from noise, for example, removing stop words.
  Fig. 1 shows the result after removing stop words from a fragment of educational content.

4. Document vectorization (for example, using the TF-IDF method):

- TF-IDF takes into the account the frequency of occurrence of each word in the document and the inverse frequency of the text.
- Each fragment of educational content (question, task, answer, requirements for mastering, recommendations, etc.) is represented as a vector, where each component corresponds to the TF-IDF value for a specific word (term, concept) in the corresponding fragment (topic, test, task, etc.) of the natural language text.

Text fragment
Artificial intelligence provides unique capabilities that seemed like fantasy a few years ago. Algorithms can recognize voices, identify people, detect errors, give recommendations, decode emotions, etc. Artificial intelligence successfully performs various monotonous operations and processes large volumes of data.
Removing stop words
[‘Artificial’, ‘intelligence’, ‘unique’, ‘opportunities’, ‘fantasy’] [‘Algorithms’, ‘voice’, ‘people’, ‘errors’, ‘recommendations’, ‘decode’, ‘emotions’] [‘Artificial’, ‘intelligence’, ‘monotonous’, ‘operations’, ‘processes’, ‘large’, ‘volumes’, ‘data’,]

Figure 2: Result of removing stop words from natural language text

5. Document vectorization (for example, using the TF-IDF method):

    - TF-IDF takes into the account the frequency of occurrence of each word in the document and the inverse frequency of the text.
    - Each fragment of educational content (question, task, answer, requirements for mastering, recommendations, etc.) is represented as a vector, where each component corresponds to the TF-IDF value for a specific word (term, concept) in the corresponding fragment (topic, test, task, etc.) of the natural language text.

6. To determine the semantics of a fragment of natural language text (including regarding students' skills and key competencies), the RAKE (*Rapid Automatic Keyword Extraction*) algorithm is used. This algorithm classifies the meanings of words and phrases based on the frequency of occurrence and the number of repetitions to create a structured list of key competencies for each fragment of educational content mastered by the student. Defining keywords makes it possible to give a concise overview of the content of large natural language texts, find fragments similar in keywords and create appropriate semantic connections. For each line, there is a list of keywords that can be used to analyze and understand the semantics of each fragment of educational content. The result depends on the properties of the text and the reaction of the RAKE algorithm to specific content.
7. Formation of recommendations that will take into the account the features of the student's mastery of educational content. Among the approaches to forming recommendations, the following should be noted:

    - Use of collaborative filtering, which is based on the analysis of the advantages and results of mastering educational content by other students similar in profile to the current student.
    - Use of content-oriented filtering, which directly analyzes the content of a fragment of educational content and their correspondence to the user's profile.

Personalized recommendations allow students to quickly find fragments of educational content that are relevant to their levels of knowledge and detail.

6. Use of ontologies in semantic analysis of natural language texts in education

The development of information technologies has led to the emergence of the so-called Smart-education, which is based on the ideas of:

- Individualization of training (learning).
- Involving students in professional activities at an early stage of training.
- Increasing students' motivation for learning (including professional, developmental—based on self-education).

Let us consider the methodology for forming knowledge components of educational content based on ontology, their use for designing training courses.

Innovative activities in training should be aimed at the use of learning (educational, training) and information technologies, within the framework of a single paradigm of education—a basic model of a specific way of organizing educational information based on the properties of generality and variability.

The proposed methodology for forming knowledge components of educational content is based on the concepts of:

- Software engineering and knowledge formalization.
- Ontological engineering for the representation and organization of semantic knowledge of educational resources (educational content).
- Construction of system abstractions of educational content based on the properties of commonality and variability, which makes such content flexible and adaptive to changing modern requirements and learning conditions.

Let us consider the representation of educational content using an ontology and a characteristic model.

Ontology defines the conceptualization that underlies the formalism of the knowledge representation [5, 6].

Modeling of characteristics is the main technique for identifying and recording commonality and variability in concepts in the ontology and in the properties of characteristics, which allows developing reusable educational components and applying them to design training (learning) courses.

The use of ontology and characteristic models in the learning process requires:

- Analyzing the structure and organization of educational content.
- Creating images using associative linking of concepts into structural elements of educational content, allowing the formation of a holistic system of knowledge of a separate course and specialty.
- Including mechanisms for the influence of educational content images on students, contributing to an increase in their cognitive ability and the acquisition of professional competencies.

The construction of an ontology begins with identifying the basic concepts (terms, concepts) of the educational content, the set of which determines the semantic knowledge of this content.

By the ontology of a basic concept we mean a hierarchical structure of concretizing concepts connected by the relations:

- “Composition”
- “Aggregation”
- “Alternative”.

With the help of the ontology of a basic concept, knowledge is described.

The characteristic model implements the configuration aspect of the ontology due to the properties of commonality and variability of the concepts of the ontology.

The commonality of requirements for individual fragments of educational content within the framework of a training/learning course (courses) determines the similarity of their characteristics, which allows adapting the educational content to modern requirements and professional competencies of students.

Reusable knowledge components should differ from traditional educational content (educational resources) in their dynamism and variability, the main method for identifying which is the modeling of characteristics.

Characteristics are indispensable, for example, in a brief description of educational content.

Characteristic models allow for formalized modeling and presentation of the semantic content of educational content.

The most commonly used knowledge representation models include production, network, frame, algebraic models, graphs and sets.

In the artificial intelligence, knowledge about the problem subject area is represented as a hierarchy of structured objects linked by relationships.

This idea underlies such knowledge representation formalisms as:

- Frames
- Semantic networks
- ONTOLOGIES
- the UML language, which (being a language for representing knowledge in the form of a hierarchy of structured classes) allows describing declarative knowledge of the problem subject area.

The rules for representing knowledge are based, in particular, on the fact that [1, 8, 16]:

1. Semantic knowledge of educational content can be represented by a set of supporting concepts, each of which is identified by its subconcepts (the so-called "daughter" ("child") concepts).
2. By a concept we mean any thought that reflects the main properties and relationships of objects (objects, phenomena, processes) of the educational content.
   With the help of concepts, knowledge is systematized. Concepts are subjective, since their semantics are determined by the context of application.
3. Ontology is a detailed specification of the conceptual structure of educational content. Ontology allows one to define the formal semantics of some knowledge.

The development of an ontology is necessary, in particular, when:

- Sharing a common understanding of the structure of concepts of educational content.
- Modeling concepts of educational content, which requires an analysis of the correspondence between the object and its properties and for perceiving the object as a variant of the concept.
- Designing system abstractions of educational content based on the properties of commonality and variability.
- Reuse of knowledge in the design of information training systems (or educational programs of specialties).

Ontology is defined as a triple [5, 6]:

$$O_m = < C, R, F >,$$

where *C* is a set of concepts (terms) of educational content; R is a set of relations between concepts; F is a set of interpretation functions, the definitions of which are specified on the relations between concepts in the ontology.

Parent concept is an abstract component expressing commonality for all its "child" concepts.

By parent concept we mean the supporting (main, main) concept of educational content.

An instance of a parent concept is considered to be a finite set of concretizing concepts of the ontology, connected with each other by the relations:

- "Composition"
- "Aggregation"
- "Alternative"

with the help of which the semantic identity of each of the concretizing concepts with its parent concept is realized.

Visually, an ontology is represented by a directed graph (ontograph) *G*, the vertices of which are concepts, and the edges are the relations between them.

Ontology is the embodiment of conceptual knowledge about the problem subject area.

It consists of the following structural components:

- Taxonomy
- Descriptions of the relations in which the problem subject area objects are located.

Creating ontologies is a complex and iterative process. It involves experts in specific problem subject areas and knowledge engineering specialists.

To date, approaches have been developed that allow this process to be automated to a certain extent.

However, the vast majority of existing ontologies have been developed "manually" using special technical tools—ontology editors (for example, Protégé, OntoEdit) [23, 24].

Ontologies are often developed as part of solving one specific problem, and the requirements for the ontology are dictated by the specifics of the chosen approach and the goals set.

Ontologies are not created once "for centuries". During the life cycle, they can change significantly because:

- Ontologies can contain errors embedded at the design stage.
- Concepts about the problem subject area can change over time, which will make a number of assumptions irrelevant or contradictory to reality.
- Requirements for the ontology themselves can also change over time.

The final concept is the category "Semantic Identity".

For this, we introduce the following definitions of the interpretation function on the corresponding relations between ontology concepts:

- The "Composition" relation reflects the property of commonality for the "child" concept and the mandatory presence of the child concept in all instances of the parent concept.
- The "Aggregation" relation reflects the property of commonality for the "child" concept and the optional presence of the child concept in instances of the parent concept.
- The "Alternative" relation reflects the property of variability (dynamics) of the "child" concept and the optionality of its presence in instances of the parent concept.

To display knowledge, an algebraic model of knowledge is adopted, which is presented in the form of a knowledge expression—a specially developed notation representing a sequence of concepts and operations on them, with the help of which the supporting concept of the ontology is identified.

For example, in the knowledge expression:

$$Con_i <= {}^*Con_{i1}({}^*Con_1 \sim + Con_2) + Con_{i2},$$

the supporting concept Con_i is specified:

- By the mandatory concept Con_{i1}, which is specified by the following "child" concepts:
  - By the mandatory concept Con1
  - By the optional concept Con_2
- By the optional Con_{i2}.

The operator < = denotes the "Implication" relation, i.e. implication is associated with causality.

The description of the ontology of the problem is performed in the OWL language [25].

7. Algorithmization of automatic text processing

A feature of the development of linguistics in our time is a close relationship with NLP [1, 26, 27].

Within the framework of NLP, algorithms for processing units of natural language (linguistic algorithms) are developed and applied, which can be classified taking into the account the following criteria:

- Communication method
- Speech form
- Level of intelligence
- Level of the language system.

Linguistic algorithms for text analysis are widely used in:

- Information retrieval systems
- Automatic abstracting systems
- Information training systems (if they involve analysis of natural language texts of corresponding fragments of educational content).

Among the algorithms for linguistic text analysis, the following should be highlighted:

- Algorithms for processing monologue speech (mainly the texts of scientific papers).
- Algorithms for processing dialogic speech (thanks to the Internet, reflected in chats, blogs, forums).

According to the level of intelligence, algorithms developed for intellectual analysis of text (text mining) can be distinguished.

As a result of applying these algorithms, the most significant information contained in the text is revealed.

Morphological analysis algorithms allow recognizing elements of the word structure—roots, base, affixes, endings.

Such algorithms include stemming and lemmatization.

The purpose of stemming is to identify the bases of semantically similar word forms.

This is necessary for adequately weighing the terms presented in the texts (for example, fragments of educational content) in order to facilitate the process of information retrieval.

A stemmer processes the text into a list of word bases for this text. Stemmers can be algorithmic and dictionary-based.

Algorithmic stemmers use lists of suffixes and inflections.

During morphological analysis, suffixes and endings of words in the input natural language text and in the corresponding list are compared, with the analysis starting from the last symbol of the word.

Dictionary stemmers use word stem dictionaries. Morphological analysis compares word stems in the input text and in the corresponding dictionary, starting with the first character of the word.

Dictionary stemmers provide greater search accuracy, while algorithmic stemmers provide greater completeness, allowing more errors that manifest themselves in:

- Insufficient stemming, when words with the same semantics are not identified by one stem; for example, the Lancaster stemmer identifies childr—as the stem children; in this case, the stem childr cannot be used to identify the plural (children) and singular (child) of one lexeme.
- Excessive stemming, when words with different semantics are identified by one stem; for example, the Lancaster stemmer identifies bet as the stem *better*; in this case, based on the *bet* base, the adjective *better* is identified with the verb *bet* and its derivatives (*bets*, *betting*), the meaning of which has nothing in common with the meaning of the adjective.

Algorithmic stemmers are more common than dictionary stemmers.

This is explained by the fact that the number of suffixes and inflections in each language is small.

Therefore, changes at the level of morphological structure occur more slowly than at the lexical level.

Rapid social and technological development causes the disappearance of some words in speech and the appearance of others.

The large size of the dictionary (when using dictionary stemmers) also reduces the speed of the system. The most famous algorithmic stemmers for the English language are: Porter's stemmer and Lancaster stemmer [28].

Y-stemmer performs morphological analysis based on annotation with parts of speech tags.

This allows you to take into the account only suffixes and endings that correlate with the part of speech for a given word.

In Y-stemmer, irregular forms of verbs, nouns and pronouns that form plurals irregularly.

The effectiveness of morphological analysis is determined by the concept of stemmer power, which is measured by:

- The ratio of the number of word forms in the original text and the word stems remaining after stemming.
- The number of characters contained in the removed suffixes and endings.

Lemmatization involves identifying word stems, taking into the account the parts of speech to which the word forms belong.

A stemmer will identify read, reads, reader, readers with one stem read, while a lemmatizer will identify the verb forms read, reads with the stem read, and the nominal forms reader, readers with the lexeme reader.

A lemma is a lexeme, the task of lemmatization is to identify word forms that are related to one lexeme.

Lexical analysis algorithms recognize lexical units of a natural language text.

One of the algorithms of lexical analysis is lexical decomposition, which involves breaking the text into tokens using programs called tokenizers.

Tokens coincide with word forms.

For lexical units of text, the term "token" is used, not "word", since a token can be understood as a unit of language smaller or larger than a word.

Most tokenizers have been developed for English language.

Tokenizers perform decomposition based on spaces between words and usually recognize apostrophes and the characters following them

('s, 'll, 'd, 'm, 't, 've, 'ref)

as separate tokens; punctuation marks are separated from words and removed [29].

Recognition of phrases and abbreviations is performed using regular expressions.

Abbreviations such as *e.g.* are one token; the same goes for a date, such as *11.01.2025*, is one token.

Initials are often considered as separate tokens. This will allow for adequate weighting.

If different people are meant, then the last name and initials should be considered as one token.

Lexical decomposition is performed based on lists of abbreviations. Stemming involves first breaking the text into tokens.

Based on the list of tokens, the following is performed:

- Syntactic decomposition
- Weighting
- Annotation performed at the lexical level.

Annotation of natural language texts is carried out by special programs—taggers.

Taggers transform the list of tokens into a list in which each token is assigned a tag indicating its linguistic characteristics.

A common type of tagger is part-of-speech taggers (POS taggers), which recognize the part of speech of a token and assign it a corresponding tag.

In addition to information about the part of speech, information about the lexical, grammatical and semantic characteristics of the word is also indicated.

For example:

- NN is a common noun in the singular
- NNS is a common noun in the plural
- AJC is an adjective in the comparative degree, etc.

Lists of part-of-speech tags differ in the degree of granularity.

A more granular classification provides more information, but also causes a greater number of errors.

Part-of-speech taggers perform:

- Tokenization
- Morphological classification
- Disambiguation.

Morphological classification involves:

- Matching each token of the natural language text with the dictionary.
- Assigning tags of parts of speech to it. Many words are associated with only one part of speech (prepositions, articles, pronouns), but there are words that are used as different parts of speech.

Homonymy of verb and nominal forms is typical for the English language.

The corresponding statistical information is important at the stage of disambiguation.

If a word from the text is not in the dictionary, then the rules for recognizing the part of speech to which it belongs are applied.

For example, if a word ends in *-ious*, then it is assigned an adjective tag, since such an ending is typical for English adjectives.

Words that begin with a capital letter are assigned a proper name tag.

If it is impossible to apply the rules, then the token is assigned a noun tag, which is used by default.

Some tokens can be assigned more than one tag.

Then statistical information about them is used for disambiguation, which involves choosing one of two or more tags assigned to such a token.

Depending on the disambiguation algorithms, part-of-speech taggers are divided into stochastic and rule-based.

Stochastic taggers analyze the probabilistic parameters of each tag, and as a result, the tag with the highest probability value is selected [30].

In rule-based taggers, the frequencies of tag use with a particular token are taken into the account.

Such a tagger is trained on a large annotated corpus, memorizing the most frequent tags of morphologically homonymous word forms.

When setting up a tagger, the following groups of rules are used:

- Rules that take into the account the lexical parameters of the current token.
- Rules that take into the account the context of the token.
- Rules that take into the account the distance from the current token to another token with a certain lexical parameter.

It is advisable to train the tagger on 90% of the corpus texts; 5%—for testing and error recognition: the tagger annotation is compared with the corpus annotation; the effectiveness of the rules is assessed on another 5%.

Without applying the rules, the tagger allows about 8% of errors, and after applying and refining the rules—up to 3.5%.

Dynamic annotation is used in factographic search systems, for example, with the help of such semantic tags as:

- Person
- Location
- Course (Group)/Department
- Organization.

Annotation with tags of cognitive roles (knowledge roles) is used in text mining.

Annotation with semantic and cognitive roles involves the recognition of individual words and phrases.

Such annotation requires preliminary development and application of special grammars of phrase structure at the syntactic level of the language system.

Among the algorithms of syntactic and discourse analysis, we note syntactic decomposition (syntactic splitting).

Programs that implement these algorithms are called splitters.

Splitters convert natural language text into a list of natural language sentences.

These algorithms recognize sentences based on text formatting symbols:

- Spaces
- Punctuation marks, etc.

Splitting text into sentences is complicated, in particular:

- Due to the lack of standard text formatting.
- Periods, exclamation marks, question marks (which are usually used as separators) can be used not only at the end, but also in the middle of a sentence.

Sentences are the basic unit of analysis. Often, text units that are formatted as sentences are not actually sentences.

These include such elements as:

- Table of contents
- Headings of individual sections
- Titles of figures
- Tables
- Text used within tables and figures
- Headers
- Footers.

The deductive-inversion architecture of text decomposition assumes the following:

- The text is split into paragraphs.
- Paragraph is broken down into words.
- Sentences are generated from words.

Decomposition begins with a larger unit (paragraph), then moves on to a smaller unit (word), the —again to a larger (sentence).

Decomposition allows ignoring such text components as headings, subheadings, and tables of contents, since they are not part of paragraphs.

Syntactic decomposition is the basis for a number of algorithms for recognizing the phrasal structure of a sentence.

Such algorithms include algorithms for extracting *n*-grams—phrases consisting of two, three or more tokens.

The breakdown into phrases is carried out taking into the account the position of the token in the sentence.

Recognition of n-grams is carried out based on the corresponding rules.

Analysis of the n-gram distribution allows identifying statistically significant phrases and is often used in algorithms for annotating parts of speech with tags.

In this case, the beginning and end of a sentence are designated by some conditional tags (false tags), which allows even sentences consisting of one token to be considered as trigrams and to establish the probabilistic parameters necessary for selecting a particular tag.

N-gram distributions are used for automatic classification and categorization, since they act as an important parameter that allows determining the belonging of the text to a certain:

- Category
- Type
- Group
- Genre.

When analyzing at the syntactic level, bigrams and digrams act as the main units.

Higher-order n-gram analysis is used for:

- Automatic spelling correction
- Automatic text recognition (Optical Character Recognition), where the main units are symbols in tokens.

Chunkers are used to analyze morphologically significant phrases, which generate lists of phrases of a certain type (in particular, nominal, verbal).

The most common are noun phrase chunkers, recognizing phrases with a control noun, which acts as keywords (supporting concepts) reflecting the main content of the text.

The rules of phrase structure were developed within the framework of *N*. Chomsky's grammar concept [31].

Grammatical rules are written in the form:

NP —> NN
NP —> DetNN
NP —> DetANN,

where the composition of the phrase, in this case a noun phrase (*NP*), is indicated, as well as the word order:

- In the first case, the noun phrase consists of only one noun (NN).
- In the second case, of a determinant (Det) and a noun, with the determinant taking a position before the noun, and the reverse word order is incorrect.
- In the third case, the phrase consists of a determinant, an adjective (A), and a noun, while other word order options are incorrect.

At the syntactic level, decomposition can be carried out into:

- Phrases
- Sentences
- Clauses-elementary predicative structures expressing a judgment; clauses are distinguished by formal features, which may include, for example, the presence of a noun phrase and a verb phrase following it.

After filtering, noun phrases are identified. Groups with large weights are selected as terms.
Unique (non-repeating) noun phrases receive the largest coefficient.
When used repeatedly, the coefficients (uniqueness and all others) are reduced by half.
Such an algorithm uses grammar, with the help of which types of:

- Phrases
- Syntactic roles
- Lexical parameters
- Grammatical parameters

are recognized.

In automatic abstracting, replacing pronouns with nouns (terms, keywords, reference concepts, etc.) allows for adequate weighting of terms.

Among the algorithms for automatic text analysis, one can also distinguish:

- Surface-level algorithms (performed on the basis of dictionaries containing statistical and probabilistic data on the distribution of language units).
- Algorithms of the semantic-syntactic level (implemented on the basis of dictionaries-thesauri, semantic dictionaries, ontologies).
- Algorithms of the discursive level.

The development of algorithms for the analysis of dialogic texts is a promising direction in the automatic processing of natural language texts.

8. Analysis of natural language texts in given context

Modern natural language text analyzers are capable of providing the process of knowledge extraction from texts of educational content fragments mainly in English.

Processing of coherent natural language texts is performed by linguistic analyzers.

For processing unstructured natural language texts, methods of constructing a formal object structure are used.

Search and knowledge extraction should occur in a certain context.

The stages of the text structuring algorithm using key concepts are, in particular:

- Defining key concepts of the natural language text of educational content.
- The main semantic terms.
- Conducting a preliminary discourse analysis based on key concepts.
- Adding other concepts only if they provide a connection between key concepts.
- Consolidating (merging) intermediary concepts.

Classification using the Precision-Recall model [32] forms a number of natural language texts with a correlation coefficient of the text relative to a given discourse.

When using Precision-Recall, no attention is paid to the features of lexical and syntactic analysis of languages.

The main difficulty is that it is necessary to identify the most plausible features by which the text can be classified.

The features are organized in the following structure:

- Surface features
- Syntactic features
- Lexical features
- Reference features
- Discursive features.

Currently, combined approaches are often used using both elements of linguistics:

- Corpus analysis
- Analysis of linguistic concepts

and machine learning algorithms [33].

Network educational resources and information learning systems and/or platforms support online learning and are aimed at different categories of students.

The most effective ones are based on ontologies that describe the semantics of the presented information resources [1, 5, 6].

One of the most difficult problems for automated teaching aids is the problem of adapting learning material (educational content) to groups of students with different levels of learning.

Case method, or the method of situation analysis, is an interactive training (learning) method designed to develop the personal component of knowledge in students.

This cannot be transferred directly by the teacher to the student.

Such aspects of knowledge (basic, professional, etc.) can only be obtained by the student independently during the analysis of the presented situations, which obviously do not have a single correct solution.

Simple accumulation of natural language texts in poorly structured information repositories makes these materials practically inaccessible due to their quantity and the ineffectiveness of

searching only by keywords, when the semantics of words and the meaning of the context in which they are used are not taken into the account.

The above-mentioned problems of processing natural language texts can be solved, in particular, using the case method.

The features of this method are as follows:

- The problem of selecting material or organizing cases in a sequence, where each element carries novelty, potentially new knowledge and experience, while maintaining the thematic outline. Requirements for this sequence:
  - Consistency
  - Uniqueness
  - Increasing complexity of the educational content.
- Sequential mastering of cases (composed of natural language texts of fragments of educational content) taking into the account previously acquired knowledge and competencies makes the analysis of current fragments of educational content more effective and productive.
- Within the framework of this problem of thematic grouping of natural language texts, the issue of adapting educational content to students with different levels of training (both basic and professional) is resolved.
- Due to the obsolescence of natural language texts of fragments of educational content, the effectiveness of such a case is much inferior to the planned one.

Therefore, it is necessary to select an alternative with changes to the outdated case while maintaining the general focus of the entire sequence of fragments of educational content (or educational content of the entire course).

An information learning system requires working with text fragments of educational content at the level of word semantics and the context in which they are used.

Solving the problem of semantic processing of natural language texts is closely related to the issues of representing knowledge of the problem area and methods of annotating texts.

Knowledge of the problem area allows one to operate with the text as a set of objects and facts identified and interpreted within the framework of existing knowledge.

Type systems and ontologies were considered in the study (on semantic text processing based on the considered algorithms) independently of each other.

For example, ontologies serve (mainly) to represent conceptual knowledge of the subject area and logical inference on them, and word type systems: word forms, lexemes, terms (concepts), and fragments of natural language text, etc. should be developed to analyze a specific natural language text.

However, given the complexity of developing ontologies and taxonomies, it makes sense to form type systems directly from the description of ontologies.

A system built using ontologies allows optimizing the solution, in particular, of the following tasks:

- Simplifying the process of creating a type system used at the stage of analyzing natural language texts.
- The entire process comes down to developing an ontology of the problem area.
  At this stage, specialized visual tools (for example, Protégé [23]) can be used.
- Using ontologies that are constantly modified and use the current type system.
- This approach will avoid re-analyzing the entire set of fragments of educational content.

Conclusions

The proposed approach forms the basis of a system for semantic processing of natural language texts (training cases, fragments of educational content). The use of methods for semantic analysis of natural language texts is more effective than classical approaches to processing unstructured information. The use of ontologies allows extensive experience in the field of knowledge representation about the corresponding problem areas of the relevant learning (training) courses.

The proposed approach to the analysis of natural language texts of the educational content of an information learning system, based on ontology, guarantees that, when annotating fragments of educational content, the current version of the type system reflecting the essence of the problem is always used and the database of annotated fragments of educational content is always up-to-date and does not contain the results of linguistic analysis of these fragments (especially natural language texts formulated as answers to questions and/or results of completing individual/independent assignments).

The ontological approach to presenting educational content in natural language promotes to creation of opportunities for the so-called "free" communication between the users of the information learning system during educational/learning/training processes, which, in particular, helps to increase the efficiency, individualization and quality of educational/learning/training processes, increase the level of students' motivation for education (learning), increase the volume of educational content of the relevant course (learning topic), and implement self-testing of acquired knowledge, competences, etc.

Declaration on Generative AI

While preparing this work, the authors used the AI programs Grammarly Pro to correct text grammar and Strike Plagiarism to search for possible plagiarism. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] K. O. Tkachenko, Using of NLP methods in intelligent educational systems, digital platform: Inf. Technol. Sociocult. Sphere 7(1) (2024) 80–96. doi:10.31866/2617-796X.7.1.2024.307009

[2] One-hot encoding in NLP. https://www.geeksforgeeks.org/one-hot-encoding-in-nlp/

[3] P. S. Reddy, et al., A study on fake news detection using naive bayes, SVM, Neural Networks and LSTM, J. Adv. Res. Dyn. Control Syst. 1(30) (2019) 942–947.

[4] J. Cheng, L. Dong, M. Lapata. Long short-term memory-networks for machine reading, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Stroudsburg: Association for Computational Linguistics, 2016, 551–561.

[5] O. Tkachenko, K. Tkachenko, O. Tkachenko, Designing intelligent multi-agent ontology-based training systems: The case of state university of infrastructure and technology, advances in computer science for engineering and manufacturing, ISEM 2021, Lecture Notes in Networks and Systems, vol. 463, 2022, 181–192. doi:10.1007/978-3-031-03877-8_16

[6] V. Pleskach, et al., Using ontologies and knowledge graphs to individualize in e-learning system, in: International Conference Information Technology and Implementation (IT&I-2023), 2023, 106–115.

[7] A. Gelfert, The ontology of models, Springer Handbook of Model-Based Science: Springer Handbooks, 2017.

[8] S. Nirenburg, V. Raskin, Ontological Semantics, 2001.

[9] Apache UIMA. URL: https://uima.apache.org

[10] R. C. Gonzalez, M. G. Thomason, Tree grammars and their application to pattern recognition, Tech. Rep. TR-EE/CS-74-10, 1974.

[11] G. Salton, Another look at automatic text-retrieval systems, Commun. ACM (7) (2000) 648–656.

[12] B. Moroz, et al., Text document classification system with Big Data technologies usage, Inf. Technol. Comput. Sci. Softw. Eng. Cyber Secur. (2) (2023) 34–40. doi:10.32782/IT/2023-2-4
[13] I. Pintye, et al., Big data and machine learning framework for clouds and its usage for text classification, Human Oriented Solut. Intell. Anal. Multimed. Commun. Syst. 33(19) (2020). doi:10.1002/cpe.6164
[14] J. F. Sowa, Building, Sharing and merging ontologies, 2009. http://www.jfsowa.com/ontology/ontoshar.htm
[15] P. Jain, S. Swati, A. Puneet Kumar, Classifying fake news detection using SVM, Naive Bayes and LSTM, in: 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2022. doi:10.1109/Confluence52989.2022.9734129
[16] B. Stecanella, Understanding TF-ID: A simple introduction. Monkey Learn. https://monkeylearn.com/blog/what-is-tf-idf/
[17] B. Kubekov, A. Utegenova, V. Naumenko, Applying of ontological engineering to represent knowledge and training sessions, in: 10th International Conference on Application of Information and Communication Technologies (AICT 2016), 2016 115–118.
[18] J. Pennington, R. Socher, C. D. Manning, GloVe: Global vectors for word representation, in: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, 1532–1543. doi:10.3115/v1/D14-1162
[19] Long short-term memory network, 2021. https://www.sciencedirect.com/topics/computer-science/long-short-term-memory-network
[20] D. Mhamdi, et al., Recommendation based on recurrent neural network approach. https://www.sciencedirect.com/science/article/pii/S1877050923006804
[21] Transformer: A novel neural network architecture for language understanding. https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/
[22] J. Uday, J. Daksh, R. V. Aditya, A deep learning approach to job recommendation analysis with NLP, Int. J. Innov. Sci. Res. Technol. 8(11) (2023) 586–593.
[23] The protégé ontology editor and knowledge acquisition system. http://protege.stanford.edu
[24] OntoEdit, ontology engineering environment. http://www.ontoknowledge.org/tools/ontoedit.shtml
[25] Word Wide Web Consortium (W3C), OWL. Web Ontology Language. http://www.w3.org/TR/owl-ref
[26] D. Jurafsky, J. H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, London: Pearson Education, 2009.
[27] Skip-Gram: NLP context words prediction algorithm, 2019. https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c
[28] M. Elias Polus, T. Abbas, Development for performance of Porter stemmer algorithm, Eastern-Eur. J. Enterprise Technol. 1(2(109)) (2021) 6–13. doi:10.15587/1729-4061.2021.225362
[29] Tokenizer: Opennlp. http://sourceforge.net/apps/mediawiki/opennlp/index.php?title=Tokenizer
[30] Y. Tsuruoka, J. Tsujii, Bidirectional inference with the easiest-first strategy for tagging sequence data, in: International Conference HLT/EMNLP-2005, 2005, 467–474.
[31] D. Pankaew, Noam Chomsky's theory of language acquisition, 2024. https://www.listening.com/blog/noam-chomskys-theory-of-language-acquisition/
[32] M. Stede, Local coherence analysis in a multi-level approach to automatic text analysis, J. Lang. Technol. Comput. Linguist. 23(2) (2008) 1–18. doi:10.21248/jlcl.23.2008.104
[33] GATE: A full-lifecycle open source solution for text processing. http://gate.ac.uk/overview.html