УДК 004.912

THE APPLICATION OF ARTIFICIAL INTELLIGENCE MODELS TO INFORMATION CONTENT COMPARISON

O. I. Komarnytska
National Academy of the State Border Guard Service of Ukraine,
46, Shevchenko St., Khmelnytskyi, Ukraine

Models of artificial intelligence to compare textual information by content, applicable on the stages of semantic and pragmatic analysis have been elaborated. The basic "content" of the answer is presented in the form of a semantic network. In contrast to the known methods of semantic and pragmatic analysis the developed algorithms based on artificial intelligence models will provide opportunities to check an automated mode text responses given in a free-text form in a natural language with greater certainty.

Key words*: analysis, algorithm, grammar, lexical unit, morphology, neural network, pragmatics, natural language, word, syntax, semantics, text.*

Problem definition. The process of assessing the knowledge and skills of students is complex and does not always ensure its validity. Till recently test control systems have been used for the automation of both current and final control of the knowledge. However, functionality of the existing systems significantly limits the ability of informal construction of the test problems. The simplicity of the software implementation has led to the fact that modern software testing tools provide the building of only certain question types. Virtually all modern systems have implemented questions that use a choice of one or several fixed options. However, for a more objective apraisal of knowledge it is appropriate to use the type of questions that provides textual answer in natural language. However, this approach is almost never used in the current testing system, as algorithm of mechanical verification of answers to natural language is quite challenging.

In order to automate the verification of answers that are given in natural language text format, one need to develop a methodology for comparing this response with the sample (samples) of the correct answer. Traditional text comparison for this task is complicated by the use of unclear, incomplete and inaccurate source data. The need for fuzzy text comparison demands taking into account (offset) ambiguity, which is inherent in the human brain and natural language. This lack of clarity, is particularly evident in the fact that the same expression can be represented as text by a man differently. And all these representations in terms of their content, are identical.

A possible promising direction for further improvement of test control may be the use of models and methods of artificial intelligence. When using a natural language an expression identical in content can be described in various ways. The structure and elements (individual words) of text representations may differ significantly from the sample. In this case, using the approach described above will be incorrect. Therefore, to compare the content of textual answers with the sample it is necessary to highlight this "content" (knowledge) as the response and a sample and compare them. The solution to this problem becomes possible using the methods of artificial intelligence.

The aim of the research is to develop the models of artificial intelligence for textual information comparison by content and use them on the stages of semantic and pragmatic analysis.

Analysis of the recent research and publications. The issue of automated knowledge control, semantic analysis of the text information has been considered by a wide range of professionals and scientists, including Askerov E. M, Badorina L. M., Hrygorova A. A., Yemelyn M. A., Yermakov A. E., Kuznetsov D. M., Radvanskaya L. N., Rudynskyy Y. D., Sokolova N. A., Stroilov N. A., Kharlamov A. A., Sharov D. A., Shydlo G. M.

Unfortunately, the analysis of the current software in this area shows that nowadays even for English there are no programs based on the use of artificial intelligence methods which are able to handle to a full extend and in a non-trivial way (for comparison) elements of "knowledge" obtained from the text [1]. This situation is due to two reasons [1]. The first is a slight spreading of linguistic analysis of text that can interpret the relation between words and, consequently, really produce knowledge as certain elements of the internal structure and suitable for content processing by "artificial intelligence". Such systems only began to appear (Net Owl (www.netowl.com), Attensity (www.attensity.com), RCO Fact Extractor (www.rco.ru)) and they have not yet integrated into applications. The second reason is the low reliability of the automatically received "knowledge" from the text. This is due to the imperfection of the modern algorithms for interpretation of the text and, in some cases, low quality sources.

The main material of the research. Semantic analysis can be used to "allocate" the content within the textual information. It allows to allocate the content structure (knowledge) from the free text in natural language. This results in identifying the content of the sentences, or their separate parts. Of course, semantic analysis is a difficult task. It is difficult to formalize. Each language has its own characteristics, which should be considered when conducting semantic analysis.

Thesaurus of the language is typically used to determine the semantic relationships between individual words. The given thesaurus is, obviously, specific to each language. It defines a set of binary relations on the set of words of a natural language. The problem of creating high-quality thesaurus is a key one when using algorithmic approach. Although, there are new commercial products that use the thesaurus in different languages, they actually include only a subset of them (English, Russian).

Therefore, the analysis of the text automatically results in the extraction of the information ("knowledge") as a network of basic concepts and relations of the weight coefficients. As a meaningful "portrait" of the text during further comparison, not merely a list of keywords is considered, but a network of concepts which, by implication, is a "reflection" of the text content. Each concept has a weight that reflects the importance of this concept in the text. Relations between concepts also have weight.

Comparison of semantic networks of the two texts will enable a comparison of their content. Regardless of the construction of the sentences, the presence of the additional judgments, minor quality characteristics that can be present in the response, the main "content" in the form of a semantic network can be defined in the text. A similar procedure is carried out with the "model" of a correct answer. Comparison of two semantic networks (text and sample answers) allows to estimate reliably the degree of identity and as a result to put a valid mark.

In the previous studies [2-4], there was proposed an algorithmic implementation of the verification of textual answers that was used in the testing system. Checking algorithm was based on fuzzy text responses compared with samples of the correct answer. When conducting a comparison not only the presence of words but also their order in the text was taken into account. Thus, a sentence structure was considered to some extent.

However, when using natural language the same in content utterance can be described in various ways. Of course the structure of the text representations may significantly differ from the sample. In this case, it will be incorrect to use the approach described above.

Thus, in order to compare the content of textual answers with the sample, it is necessary to define this "content" (knowledge) both from the response and the pattern and make a comparison. Effective resolving of this problem is possible by using methods of artificial intelligence.

In modern science, there are two approaches to understanding of the term "knowledge" in the context of its isolation from the text. Within the first approach, the attention is focused on the pragmatic aspect (used in the direction of «knowledge management»), when knowledge is represented as data received in a certain place at a certain time and needed to

solve a practical problem, for decision making, and so on. In this case, neither the structure nor the mode of representation of the "knowledge" differ from the other data. They may be presented by a fragment of a text document, picture, part of the database, etc. Within another approach, an emphasis is placed on the content aspects, and when viewed within the "artificial intelligence", it implies that knowledge differs from normal data by its structure.

Various information about the world can be represented in the text form. However, the ambiguity of the representation, which is inherent to almost all natural languages (including Ukrainian), makes the task of content selection extremely difficult. Despite over 50 years of research in artificial intelligence, there are still no universal solutions to most applied word processing problems. This is due to the problems of formalization of natural language. Natural language is a very complex semiotic system that consists of an unlimited number of subsystems, each of which is finite and, therefore, can be formalized. However, the language itself is an open system that cannot be fully formalized [1].

By processing the text, it is possible to identify the following stages of analysis: morphemic, morphologic, syntactic, and semantic.

The latter type of analysis, semantic, enables allocation of the content structure (knowledge) from the free-text natural language. At this stage an identification of the content of the sentences or their separate parts takes place. Among all steps of text information processing, semantic analysis is the most difficult. It is difficult to formalize. There is no doubt that every language has its own characteristics to be considered when conducting semantic analysis.

Universal approach is also possible. It is based on the use of neural networks. Neural network can be taught to compare text information in different languages, as this is done by its biological prototype - human brain. During the internal hidden processes that occur when using neural network, an information preprocessing (morphological, syntactic analysis) and semantic analysis are carried out. However, to enable solving of complex problems such as semantic analysis, neural network must have a large enough capacity. The learning procedure is a problematic one. The weights are formed in it and they define the memory of the neural network. It is necessary to build a significant number of phrases to be compared for teaching just one subject area. These phrases should include all terms used in this subject area. They should cover all the concepts that must be preset in the memory of a neural network. For a more accurate memorizing in the process of comparison, one should select not only essentially different text pairs, but also similar in spelling and different in meaning.

The overall structure of the neural network is shown in the picture. The first, open layer of a neural network, is represented by two groups of neurons *1N*. Login information is entered in this layer. In the first group of neurons, there is the text information which is verified with the sample, in the second group of neurons – the sample. Since neurons use single-bit coding information, and one character in many encodings is represented by one byte to store information about a single character, it requires 8 neurons of the input layer. Thus $N=8\,L$, where L - maximum length of sentences in characters.

For the primary information processing in neural networks, a layer of neural network with two groups of neurons *1..M* is used. Between neurons of the input layer and the first layer of each of the two groups, there are established connections. To improve the quality of primary processing (morphological, syntactic analysis), it is possible to include another similar layer of neurons to the neural network.

Basic information processing is implemented in the next hidden layer of neurons *1..K*. Each neuron in this layer enters information on each of two groups of neurons of the previous layer. In order to increase the information capacity of the neural network by improving the capabilities of semantic analysis, it is possible to increase the number of neurons in this layer (*K*) and to include the following additional intermediate layers.

In order to receive the result of the test, the final layer, which consists of a single neuron, is used. One of its two possible states corresponds to the identity of the two texts in content. Another condition indicates that the tests within the content differ.

The advantage of using neural networks for solving the problem of comparing text by content is its versatility. The neural network with permanent structure can be adapted (trained) to compare texts in different languages from different subject areas. However, the complexity of learning is a significant disadvantage. Moreover, studies should be conducted in all languages from all subject areas in which the neural network will be used.

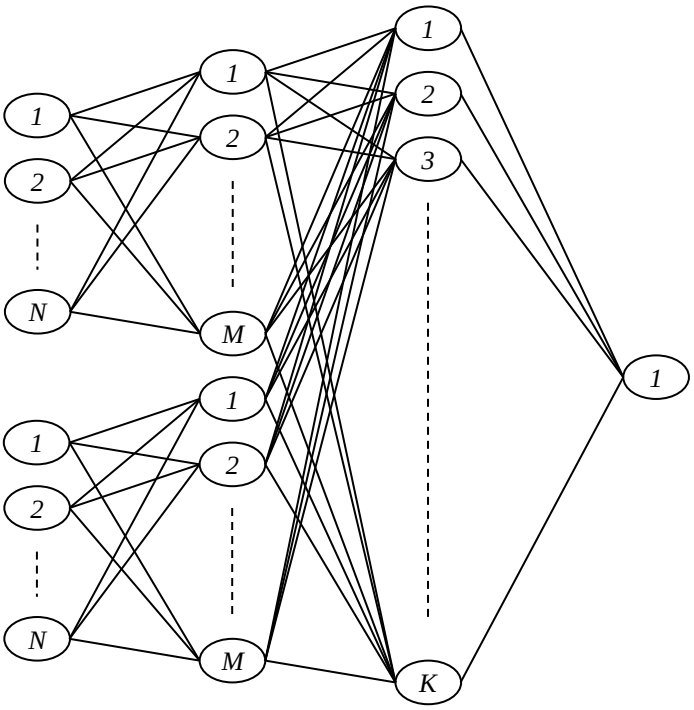

Figure - The general structure of a neural network for textual information comparison

In the absence of a hardware implementation, it is possible to simulate the neural network, but in this case, the performance will be significantly lower. It is possible to increase the speed of neural networks due to efficient use of computing resources of a computer (multiprocessing) between which a load balancing is performed. Using CUDA technology from Nvidia, the speed of neural networks may be significantly increased. It allows to use the processing power of the graphics card, GPU, which contains dozens of computational cores, to process the applied problem.

The second area of semantic analysis - an algorithmic approach. When using it on the stage of semantic text processing by algorithmic means, a detection of the content of the sentences or their parts is performed.

The statistical analysis, as well as the semantic analysis, as one of the algorithmic approaches may be used to detect key words of the text. They reflect the specificity of the text, enable one to determine its theme. In order to define them, one must take into account the semantic proximity of words by their meanings. The solution to this problem allows to build the scale of distances between words in the text. An algorithm of E. L. Ginzburg may be used for its construction [5].

To solve the problem of the semantic analysis, it is necessary to:

– develop the methods of processing the original text and making the frequency vocabulary;

– develop a filter to screen out the words that are specific to natural language as a whole;

– automate the selection of the keywords.

Known algorithms for keyword selection are divided into two groups: intertext filtering (based on the use of the frequency characteristics of a certain set of texts) and internal textual filtering (using frequency information within a text).

Algorithm for detecting the context [5] may have the following stages:

– definition of an absolute and relative frequencies of each word forms;

– search for parts of the text that contain specific word;

– definition of an absolute and relative frequencies for text parts identified in the previous step;

– comparison of the relative frequencies specified in the first and the previous step. If the relative frequency determined in the first step is smaller, then the word form belongs to the semantic field of the word.

In order to create word forms, it is possible to use a morphological dictionary.

As a result of the semantic analysis, a semantic network lays the basis for a framework for knowledge representation in the form of nodes connected by arcs (links). Properties of the resulting semantic network are:

– nodes represent concepts, objects, events, states;

– arc semantic links provide relation between the nodes - the concepts (the ratio can be of different types);

– some relation between nodes are linguistic, spatial, temporal, logical, etc.;

– concepts are organized into levels according to the degree of generality.

Thus, in the result of the text analysis the information ("knowledge") as a network of basic concepts and relationships of the weights is automatically extracted. As a meaningful "portrait" of the text during the further comparison, not only a list of keywords is considered, but also the network of concepts, which is, by implication, a "reflection" of the text content. Each concept has a weight that reflects the importance of this concept in the text. Relations between concepts also have weight quotients.

Comparing semantic networks of the two texts allows for a comparison of their content. Unlike previous approaches, this comparison enables a more reliable automated inspection of the text responses given in a free-text form in natural language. Regardless of the construction of sentences, presence of the additional judgments, minor quality characteristics that can be present in the response, the main "content" in the form of semantic network is defined from it. A similar procedure is carried out with the "model" of a correct answer. Comparison of two semantic networks (text and sample answers) allows to assess reliably the degree of identity and as a result to put a valid mark.

Thus, in the test system the previous text comparison algorithm that was based on review of available keywords and their order was replaced by a more "intelligent". The new algorithm uses a semantic network building to highlight the "knowledge". Comparison of texts is conducted by corresponding semantic networks. Using the new algorithm makes it possible to improve significantly the accuracy in the system verification of the test textual answers provided in a free form. Even with significant textual discrepancies (different approach to the construction of sentences using synonyms for the basic concepts), but the convergence of the content of texts, semantic networks that correspond to them are similar, and a new "intelligent" algorithm can properly evaluate it.

Utilization of the new algorithm with elements of "artificial intelligence" in terms of building semantic networks will significantly improve testing efficiency of the system. However, the developed thesaurus /vocabulary does not still cover all possible topics that can be used in the tests (it only contains information about the relations between the

concepts of the network theory). An improvement of the thesaurus /vocabulary in order to expand the possible test topics is the goal of the future work.

Algorithms of the artificial intelligence learning of the automated knowledge control system is based on the following algorithms: back propagation; study without a teacher; training of Hopfield and Hamming's networks.

Hopfield and Hamming's networks allow easy and effective solution of the problem of an incomplete and distorted information imaging. The low network capacity (the number of stored images) is explained by the fact, that networks do not just memorize the images, but allow their generalization, for example, via Hamming's network it is possible to classify them by maximum likelihood criterion [6-8]. However, the ease of these software and hardware models' construction make these networks attractive for many applications.

Thus, the paper presents an algorithm that can be used for the evaluation of text responses during computer testing. For a fuzzy comparison of individual words in response algorithm, Lowenstein's metric is used, but to improve the efficiency of verification, additionally a structure analysis of the sentences is conducted.

Conclusions of the research and recommendations for further scientific research. Models of artificial intelligence to compare textual information in content are used on the stages of semantic and pragmatic analysis. A semantic analysis forms the basis of semantic network - a framework for knowledge representation in the form of nodes connected by arcs (links). In the pragmatic analysis, it is determined whether a response belongs to a particular subject area. These stages are proposed for implementation through the use of neural networks. The advantage of using neural networks is versatility. The permanent structure for the neural network can be adapted (trained) to compare texts from various subject areas.

In contrast to the known methods of semantic and pragmatic analysis, algorithms based on artificial intelligence models will provide more opportunities to inspect automated text responses given in the free text form in natural language with greater certainty. Regardless of the construction of sentences, presence of additional judgments, minor quality characteristics that can be present in the response and the sample, there is allocated the main "content" in the form of semantic network. Comparison of two semantic networks (text and sample answers) provide a reliable assessment of the degree of identity and, as a result, put a valid mark.

Algorithms of the artificial intelligence learning of the automated knowledge control system are based on the following algorithms: back propagation; study without a teacher; training of Hopfield and Hamming's networks.

Experimental verification showed reasonable efficiency of the proposed algorithm when checking test questions, answers to which are given in a free form.

ВИКОРИСТАННЯ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ПОРІВНЯННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ ЗА ЗМІСТОМ

О. І. Комарницька

Національна академія Державної прикордонної служби України ім. Б. Хмельницького, вул. Т. Шевченка,46, м. Хмельницький, 29003,Україна

Розроблено моделі штучного інтелекту для порівняння текстової інформації за змістом, які застосовуються на етапах семантичного та прагматичного аналізу. Основний «зміст» відповіді представлено у вигляді семантичної мережі. На відміну від відомих методів семантичного й прагматичного аналізу, розроблені алгоритми на основі моделей штучного інтелекту надаватимуть можливості з більшою достовірністю автоматизовано проводити перевірку відповідей, наданих в довільній текстовій формі на природній мові.

Ключові слова*: аналіз, алгоритм, граматика, лексична одиниця, морфологія, нейромережа, прагматика, природна мова, слово, синтаксис, семантика, текст.*

LIST OF REFERENCES

1. Ермаков А. Е. Извлечение знаний из текста и их обработка: состояние и перспективы / А. Е. Ермаков // Информационные технологии. – М. : Новые технологии, 2009. – С. 50 – 55.
2. Ваколюк Т. В. Алгоритм нечіткого семантичного порівняння текстової інформації / Т. В. Ваколюк, О. І. Комарницька // Збірник наукових праць Військового інституту Київського Національного університету ім. Т. Шевченка. – 2013. – № 39. – С. 163-168.
3. Катеринчук І. С. Інтелектуальна автоматизована система контролю знань: лінгвістична підсистема / І. С. Катеринчук, В. М. Кулик, О. І. Комарницька // Інформаційні технології в освіті: збірник наукових праць. – Херсон : Вид-во ХДУ, 2009. – Вип. 4. – С. 139-147.
4. Катеринчук І. С. Новітні інформаційні технології оцінювання знань у вищих навчальних закладах / І. С. Катеринчук, В. М. Кулик, О. І. Комарницька // Збірник наукових робіт № 51. Частина ІІ. – Хмельницький : Вид-во НАДПСУ, 2010 – С. 56-59.
5. Гинзбург Е. Л. Идеоглоссы : проблемы выявления и изучения контекста / Е. Л. Гинзбург // Семантика языковых единиц : Доклады VI международной конференции. Т. І. - М., 1998. – С. 26-28.
6. Искусственный интеллект : в 3 кн. Кн. 2 Модели и методы : справочник / под ред. Д. А. Поспелова. – М. : Радио и связь, 1990. – 304 с.
7. Искусственный интеллект : в 3 кн., Кн. 3: Программные и аппаратные средства : справочник / под ред. В. Н. Захарова, В. Ф. Хорошевского. – М. : Радио и связь, 1990. – 368 с.
8. Искусственный интеллект : справочник. Книга 1: Системы общения и экспертные системы / под редакцией Э. В. Попова. – М. : Радио и связь, 1990. – 464 с.

Received: August 29, 2014