The Specifics of Natural Language and Ways of Processing It in the Computational Linguistics

Tomasz Panczyszyn
Faculty of Arts and Humanities
King's College London
London, WC 2R 2LS, UK
Email: Tomasz.Panczyszyn@gmail.com

***Abstract*—The computational linguistics is now undoubtedly a well-developing and prospective field of study. As an intersection between linguistics as such and the computer science, it treats many problems of how to process the natural language as to make it applicable and easily transformative for the machines. The standard question we put ourselves when it comes to the interconnected areas of the natural language processing and the computer science, is whether we can teach a computer how to speak a natural language. The issue we will be thinking over in the paper, will be the way of treating the matter of standard computational linguistics problems we do encounter on a daily basis where it connects to the field of linguistics. In this paper, a novel approach to natural language processing by using the neural networks and object oriented approach is presented.**

I. Introduction

When we try to define what the interconnections between the natural language and the computer science really are we would have to take into account the fields of study both of the disciplines regard.

Linguistics, being a study of human language consists of almost seven thousand of languages there are in the world as the subjects of examination. It actually studies all the appearing linguistic aspects, that is to say: syntax, pragmatics and semantics. So a language, with all the aspects included, is being learned by us from a minimal input in our early childhood and serves us to communicate with each other more or less easily. A high variety of languages we do have nowadays causes also that there emerges some questions regarding the translation from one language into another. The problems to be sorted out in the field are pretty most frequently caused by the fact that the semantics of the language [applies also for semiotics, syntax and pragmatics] is being acquired by us, the users of the language, quite intuitively.

As the computers and machines do not function that way, we do meet problems like the one with how to automatically process the construction of the possesives pair of words, i.e. *the photos of my friends* that a user of a natural language would rather understand as pictures presenting the friends of the person who speaks than (as it could appear in the computer translation that the photos as the property of one's firend's).

A. Related works

Natural language processing binds very strongly to the subject of artificial intelligence. The idea of creating an artificial consciousness expanded stream of science fiction literature in the nineteenth century, and rapid technological development aims to realize dreams of authors this type of books. Already in the forties of the twentieth century, an artificial neural network model has been designed. The one which has numerous applications in life today. Especially, neural networks [1] are used in problems of classification ([2], [3]) and categorization of components ([4], [5]). In [6], [7], the author's shown inference and classification system based on social media.

Another group of important methods are heuristic algorithms, which were created in order to find the maximum and minimum values of optimized functions. Heuristics proved to be a good option for finding solutions, not only for the problem of searching the extremes of functions, but also in the graphics processing [8], [9]. An example of such applications is the search of important points (called key points) to 2D images [10], [7]. Moreover, in [11], [12], [13], the authors have shown that these algorithms can also be used in the construction of unique maze. A major use of heuristics is also the problem of queuing ([14], [15], [16]) eg .: in online stores where overload may occur.

Natural language processing is such an important subject that can not only afford to develop the field of artificial intelligence [17], but also help our everyday lives, eg.: lives of blind people. Natural language processing is called parsing. One of these methods is shown in [18]. In [19], the authors presented semantic parsing using paraphrasing, again in [20] shown the idea of using semantic parsing as machine translation process. In recent years, the idea of creating computer intelligence using chatbots gaining more and more interest in recent years. In such applications, an important element is the knowledge base. In [21], the authors have shown an idea of system for development a modular knowledge base. The authors of [22] presented comparison between conversations type human-human and human-computer.

In these paper, I would like to present a novel approach to find the author of a longer text with the use of methods of artificial intelligence. The proposed model has been tested and described with regard to all its advantages and disadvantages.

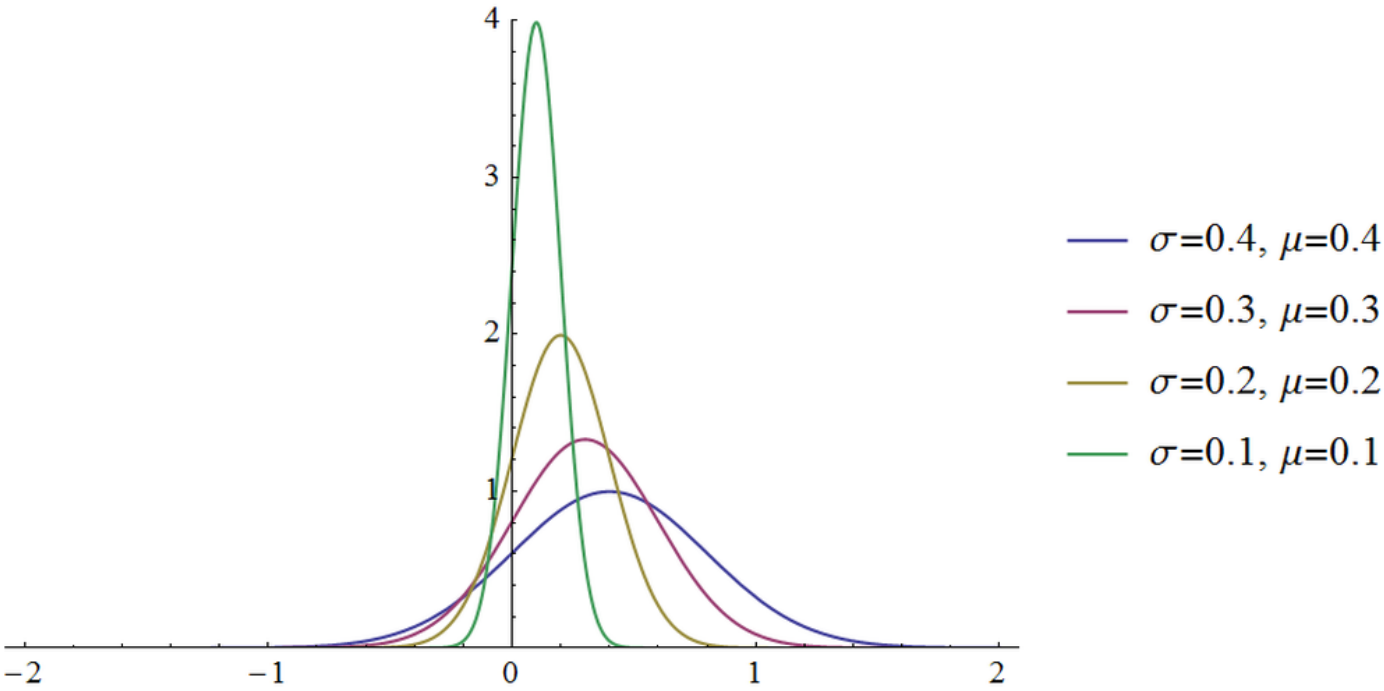

Fig. 1: Graph of a Gaussian function for different values of the parameters σ and μ.

II. Multidimensional look at modern linquistics

While studying the opportuinities of solving the problems computers do encounter when processing the languages we have to take into consideration also the relations between the words in the sentences. Even if a language is being treated intuitively, we ought to remember that the standard defining the quality of an enunciation is also the grammatical correctness. Like in the computer languages, in the natural ones as well, we have combinations that are either possible or not. Let's take an example.

We have a complex phrase

1) a) John puts on a hat every time he goes out. [John = he:possible];
   b) He puts on a hat every time John goes out [he = John: impossible].
2) a) Every time John goes out, he puts on a hat [he = John: possible];
   b) Every time he goes out, John puts on a hat. [he=John: possible].

The sentences above, therefore, show us very well that actually every native speaker of a language has naturally, with a quiet mininal imput from the childhood an intuition that tells him whether a phrase [a relation of words appearing in a sentence] could be correct and whether a grammatically and formally correct sentence could ever be constructed like this.

So what is interesting about the phrases we have been reflecting on is that there was no need of having taken syntax classes to know which of the sentences is apparently ungrammatical. Thanks to the knowledge we have nowadays, we are able to say that on the basis of the words of a natural language and the relations between them we can construct an infinite number of sentences that will be grammatically correct.

Even if we hear a sentence for the first time in our lives, we can suspect or just verify whether it is correct or not. The phrase: *The six-headed CS84 Tbs grilled the blind octopus using a MAPA mug* is surely correct when it comes to its grammar but not necessarily met by us ever before. That proves language functions intuitively.

III. Models of theoritical discription of linquistics

We can check if the order and relation between the subject and predicate is as it should be. From among many problems of either the natural language processing and the computer science issues we would like to focus just on a particular part of them. To see how many chances are given us by the modern computer sciences and its processing the language we will be trying over the article to find an answer to the question of the classification problem. The problem we would like to sort out is: given two presidential speeches from the US election can I guess with high probability whose speech it is?

Naturally, when it comes to the classification problem there is no simply answer to the question like the one above. We can, however, assume for the purposes of our study that both presidential candidates speak about recurrent themes, also probably using reccurent words. Even if there is no exact answer to this, we will try make a few assumptions and create a model to try to accomplish what we suppose.

The assumption we will be making will be based on what we know for now about the presidential speeches. We will take into consideration two presidential speeches of Barack Obama and of Mitt Romney.

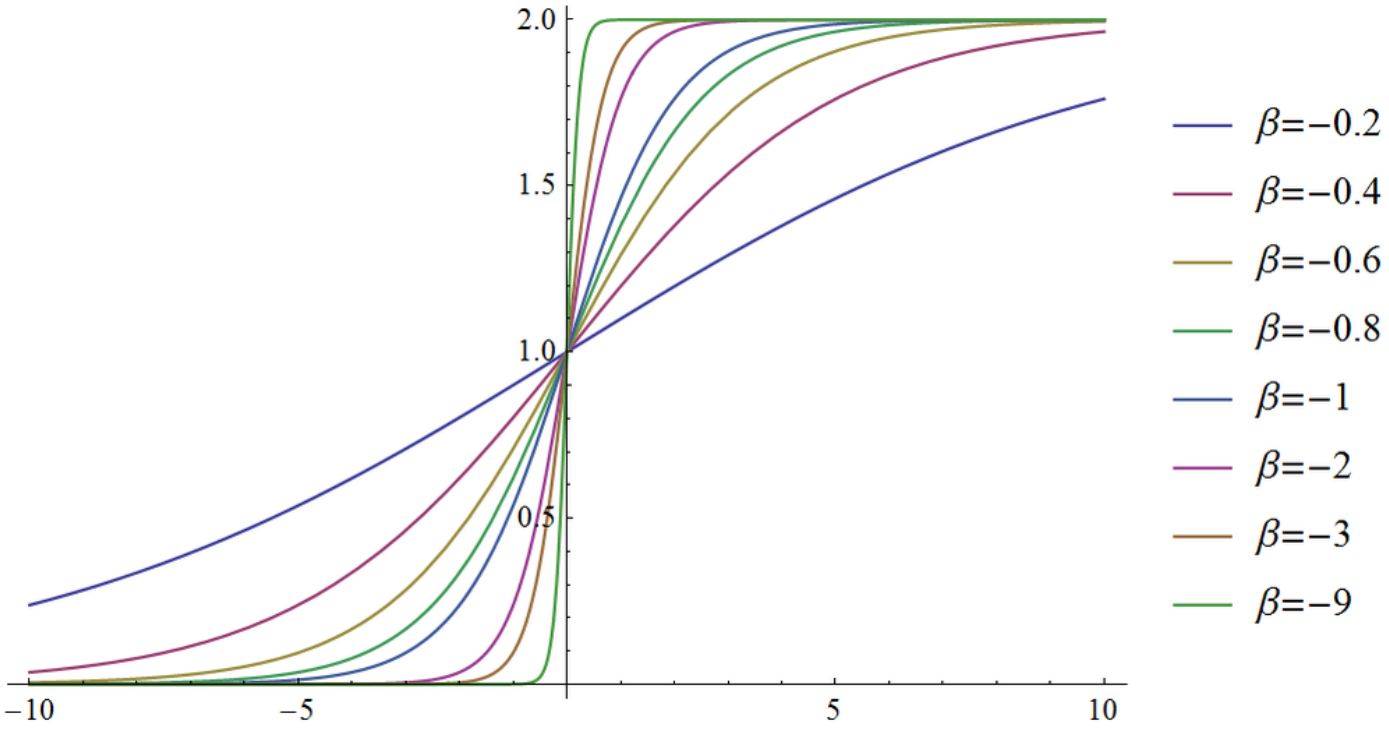

Fig. 2: Graph of an activation function for different values of the β parameters.

So just to have an idea of how it really looks like, let's have a look at the Obama's speeches on future:

"And we salute the people of Paris for insisting this crucial conference go on; an act of defiance that proves nothing will deter us from building the future we want for our children (. . .) I want to show her passionate, idealistic young generation that we care about their future. (. . .) This summer, I saw the effects of climate change firsthand in our northernmost state, Alaska, where the sea is already swallowing villages and eroding shorelines; where permafrost thaws and the tundra burns, where glaciers are melting at a pace unprecedenten in modern times. And it was a preview of one possible future."

[Barack Obama, First Session of COP 21, Nov 30, 2015].

As we have seen in the Obama's speech the word *future* appears three times. Let's now have a look at Mitt Romney's speech.

"(. . .) They came not just in pursuit of the riches of this world but for the richness of this life. Freedom. Freedom of religion. Freedom to speak their mind.. Freedom to build a life. And yes, freedom to build a business. With their own hands."

[Mitt Romney, Republican Convention, Aug 08, 2012]

The assumption we can make for now is that American people have been hearing Romney's speaking very frequently about freedom in various contexts, like: freedom of religion, freedom to speak your mind, freedom to build a business, freedom to build a life.

Followingly, what is specific for Obama's speeches is the notion of *future*. He speaks about the future of America, people's future, the better future, creating the future of dreams, future of American economy.

IV. Preprocessing of the elements of discourse

To prepare long enunciations for the purposes of an analysis of the problems of the AI, we have to represent them in the simpliest possible form, applying there some number values. To show how that actually functions we will be serving ourselves with the Bayes theorem which defines directly how the conditional probability works and can be seen as a way of understanding how the probability that a theory is true is influenced by an evidence that appears there for the first time. Let's illustrate it with a theorem

$$\begin{aligned} \text{argmax} Pr(c|w) = &= \text{argmax} \frac{Pr(w|c)Pr(c)}{Pr(w)} = \\ &= \text{argmax} Pr(w|c)Pr(c), \end{aligned} \tag{1}$$

where $Pr(c|w)$ stands for the probability of appearing of a particular word w in a particular class c of objects.

The Bayes theorem ([23], [24]) has helped us in defining both the prior and the posterior probability of concrete objects appearing in a concrete class of objects. For the most frequent appearing words we then calculate the values of y_i using the formula

$$y_i = \text{argmax} Pr(c_i) \prod_{j=i}^{m} Pr(x_j|c_i), \tag{2}$$

where c_i stands for the class of objects and the x_j – for the analyzed expression. In the next step we do process all the values by making use of the conception of blur. In the proposed method of processing we applied the Gaussian blur defined as

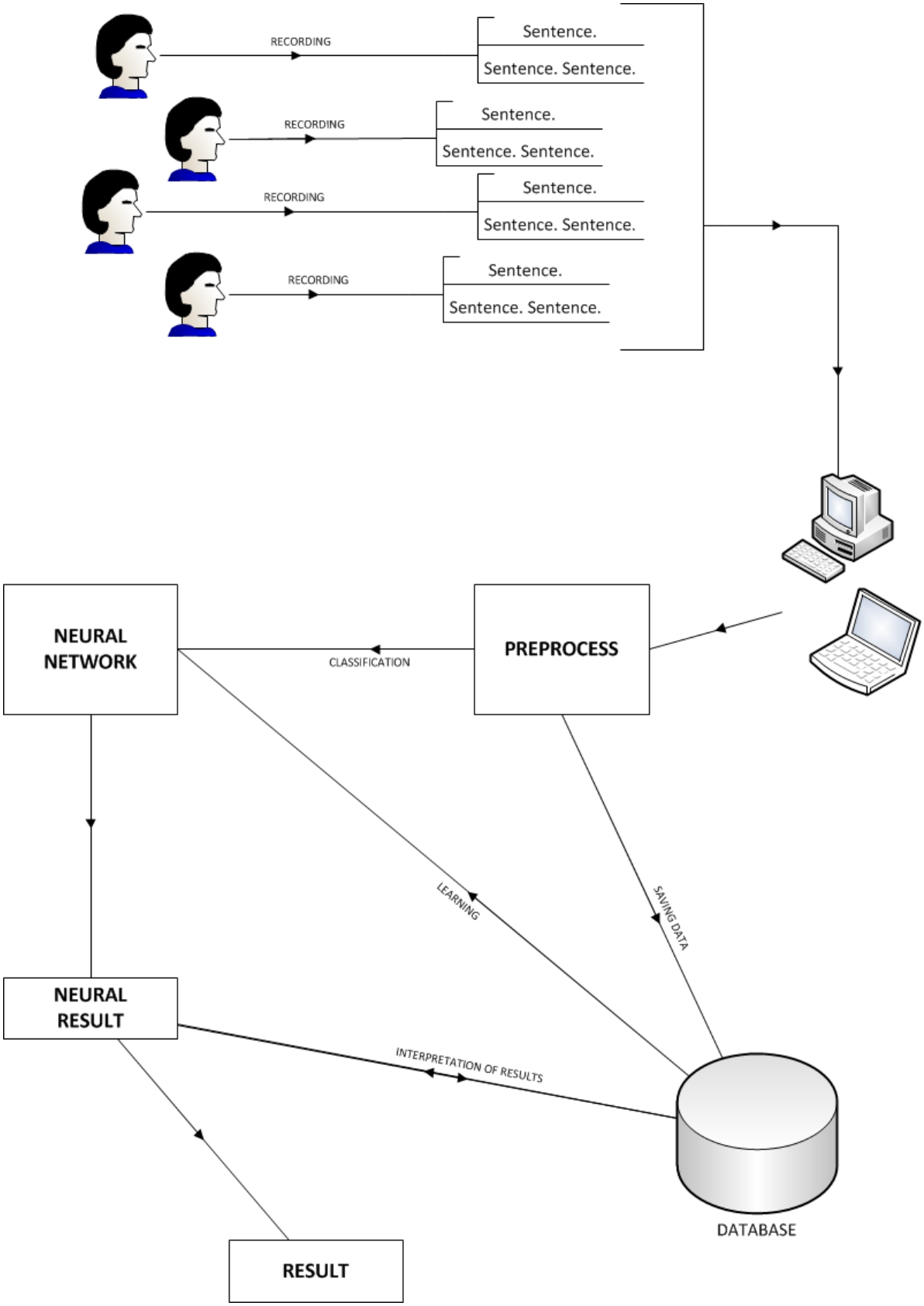

Fig. 3: The model of the proposed identification system.

$$G(x) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right), \tag{3}$$

where σ is the height of the function, and μ is the shift on y-axis.

The Gaussian blur permits us approximate the values as to assess the probability of an expression appearing in the demanded conditions. The calculated values allow us create a vector representing the enunciation desired. That could be showed by such a theorem

$$[G(y_0), G(y_1), \ldots, G(y_{n-1}), id], \tag{4}$$

where $G(y_i)$ is stand for values calculated, in accordance with the equation (3), and id means the author's numerical identifier.

V. Neural Network

Already in the forties of the twentieth century, the author of [25], [26], [27] described the first model of artificial neural network (ANN). Artificial neural network is a mathematical model inspired by the action of neurons in the human brain.

ANN is composed of three types of layers – the input, hidden and output. The input layer is responsible for the acceptance of teaching vector, and output for return a result of the network. Hidden layers are located between the input and output, they are responsible for creating a *deeper* network in order to obtain better results.

Each layer is constructed of neurons, wherein each neuron of one layer is connected to each neuron in the next layer. Neuron is the smallest object of neural network. The data enters the neuron through mergers which have a certain weight. In the neuron, all input values are rescaled by the activation function. The value after scaling is sent to other neurons along outbound connections to the next layer. As activation function selected a bipolar sigmoid function [28], [29], [30]– function is defined as

$$f(x) = \frac{2}{1 + \exp(-\beta x)}, \tag{5}$$

where β is a parameter in $(0, 1]$. Activation function graph is shown in Fig. 6.

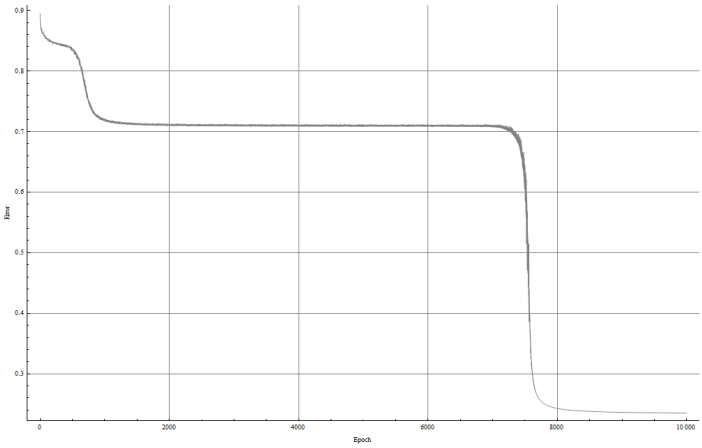

Fig. 4: Error learning neural network.

In order to increase the precision of ANN, many learning algorithms of this type of network are created. One of these algorithms is the backpropagation algorithm, which works on minimizing the error function and modify weights from the output layer to the first hidden layer [31], [32], [33]. Weights are modified using the following formula

$$w_i = w_i + \Delta w_i, \tag{6}$$

where w_i means the weight on the i-th connection, and Δ_i is calculated as

$$\Delta_i = \begin{cases} \kappa_i(1-\kappa_i)(\omega_i - \kappa_i) & \text{for output layer} \\ \kappa_i(1-\kappa_i)\sum_{j\in\kappa} w_{ji}\Delta_i & \text{for hidden layer} \end{cases}, \tag{7}$$

where κ is the output value and ω is the expected value.

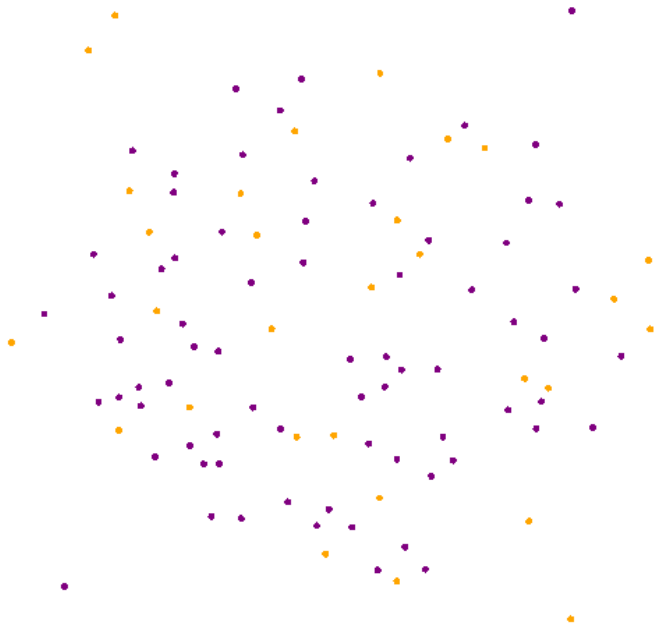

Fig. 5: Knowledge representation using Sammon mapping.

VI. Proposed model of author's identification

The problem of a person verification on the basis of a longer text requires not only the extraction of the characteristics of his speech, but also correct classification. For this purpose, the proposed model consists several stages. At the beginning, the statement is entered into a computer, where it statement is processed according to Sec. IV. The next step of action has two paths. In the first one, the sample is stored in the database. When the database contains a sufficient number of samples, the neural network is trained using the knowledge contained in the database. The second path of action is to classify the samples by the neural network.

A model of such a system is shown in Fig. 3.

VII. Experiments

The proposed solution has been tested by the use of 200 samples – 100 samples per person. Each sample contained a fragment of statements about the future, taking into account up to 60 words. For the purposes of minimizing the time of neural networks learning, each sample contained 15 components. A neural network was composed of 4 layers

- input layer composed of 15 neurons;
- 4 hidden layers composed of 4 neurons;
- output layer consisting of one neuron.

To train the network, the samples were divided into two groups (training and verifying – $80\% : 20\%$). The problem of classification has been shown in Fig. 5 using Sammon mapping – based on this interpretation of the spread of knowledge it can be seen that there can not be any easy way to separate the samples into two groups, so the problem of classification is extremely difficult. The network was trained to obtain the error of the 0.24 - error learning graph is shown in Fig. 4. In order to verify the operation of the classifier, each sample was given to the input of the network. In consequence of the operation, the system indicate the author correctly for 103 samples, which results indicates in a efficiency at about 72%.

VIII. Conclusion

The subject of the research was to prove the possibilities of processing the natural language by the methods of computational linguistics. We have shown that, given two different longer texts, we are able to identify their authors exclusively on the basis of the words used - with a minimal risk rate possible (error - 0.24). After entering the data into the computer, the use of the database's contents needed to be classified by the neural network. Having done so, we processed two independent author's speeches Barack Obama's and Mitt Romney's addresses and therefore, helped by the neural networks, we could evaluate whose a text is just on the basis of the samples given and after comparing the data we had with the one contained in the database. The outcomes then of such an experiment are of significant help when it comes to the examination of idiolects and distinctive, personal styles of constructing a discourse. It – followingly – contributes to the development of the Artificial Intelligence as it is the computer to identify the authorship of a text.

References

[1] C. Napoli and E. Tramontana, "An object-oriented neural network toolbox based on design patterns," in *Information and Software Technologies*. Springer, 2015, pp. 388–399.

[2] C.-T. Chen, K.-S. Chen, and J.-S. Lee, "The use of fully polarimetric information for the fuzzy neural classification of sar images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 9, pp. 2089–2100, 2003.

[3] A. Kandaswamy, C. S. Kumar, R. P. Ramanathan, S. Jayaraman, and N. Malmurugan, "Neural classification of lung sounds using wavelet coefficients," *Computers in Biology and Medicine*, vol. 34, no. 6, pp. 523–537, 2004.

[4] D. Valentin, H. Abdi, and A. J. OTOOLE, "Categorization and identification of human face images by neural networks: A review of the linear autoassociative and principal component approaches," *Journal of biological systems*, vol. 2, no. 03, pp. 413–429, 1994.

[5] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *arXiv preprint arXiv:1412.1058*, 2014.

[6] A. Fornaia, C. Napoli, G. Pappalardo, and E. Tramontana, "An aoprbpnn approach to infer user interests and mine contents on social media," *Intelligenza Artificiale*, vol. 9, no. 2, pp. 209–219, 2015.

[7] C. Napoli, G. Pappalardo, E. Tramontana, R. K. Nowicki, J. T. Starczewski, and M. Woźniak, "Toward work groups classification based on probabilistic neural network approach," in *Artificial Intelligence and Soft Computing*. Springer, 2015, pp. 79–89.

[8] X.-S. Yang, "Flower pollination algorithm for global optimization," in *Unconventional computation and natural computation*. Springer, 2012, pp. 240–249.

[9] M. Wozniak, "Fitness function for evolutionary computation applied in dynamic object simulation and positioning," in *Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2014 IEEE Symposium on*. IEEE, 2014, pp. 108–114.

[10] D. Połap, M. Woźniak, C. Napoli, E. Tramontana, and R. Damaševičius, "Is the colony of ants able to recognize graphic objects?" in *Information and Software Technologies*. Springer, 2015, pp. 376–387.

[11] D. Polap, M. Wozniak, C. Napoli, and E. Tramontana, "Is swarm intelligence able to create mazes?" *International Journal of Electronics and Telecommunications*, vol. 61, no. 4, pp. 305–310, 2015.

[12] D. Połap, M. Wozniak, C. Napoli, and E. Tramontana, "Real-time cloud-based game management system via cuckoo search algorithm," *International Journal of Electronics and Telecommunications*, vol. 61, no. 4, pp. 333–338, 2015.

[13] D. Połap, "Designing mazes for 2d games by artificial ant colony algorithm," *Symposium for Young Scientists in Technology, Engineering and Mathematics (SYSTEM 2015)*, pp. 63–70, 2016.

[14] M. Woźniak, W. M. Kempa, M. Gabryel, R. K. Nowicki, and Z. Shao, "On applying evolutionary computation methods to optimization of vacation cycle costs in finite-buffer queue," in *Artificial Intelligence and Soft Computing*. Springer, 2014, pp. 480–491.

[15] M. Woźniak, W. M. Kempa, M. Gabryel, and R. K. Nowicki, "A finite-buffer queue with a single vacation policy: An analytical study with evolutionary positioning," *International Journal of Applied Mathematics and Computer Science*, vol. 24, no. 4, pp. 887–900, 2014.

[16] P. V. Laxmi and K. Jyothsna, "Optimization of service rate in a discrete-time impatient customer queue using particle swarm optimization," in *Distributed Computing and Internet Technology*. Springer, 2016, pp. 38–42.

[17] X. Gao and N. Zhu, "Natural language processing," *Information Technology Journal*, vol. 12, no. 17, pp. 4256–4261, 2013.

[18] K. Xu, S. Zhang, Y. Feng, and D. Zhao, "Answering natural language questions via phrasal semantic parsing," in *Natural Language Processing and Chinese Computing*. Springer, 2014, pp. 333–344.

[19] J. Berant and P. Liang, "Semantic parsing via paraphrasing." in *ACL (1)*, 2014, pp. 1415–1425.

[20] J. Andreas, A. Vlachos, and S. Clark, "Semantic parsing as machine translation." in *ACL (2)*, 2013, pp. 47–52.

[21] G. Pilato, A. Augello, and S. Gaglio, "A modular system oriented to the design of versatile knowledge bases for chatbots," *ISRN Artificial Intelligence*, vol. 2012, 2012.

[22] J. Hill, W. R. Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations," *Computers in Human Behavior*, vol. 49, pp. 245–250, 2015.

[23] K.-R. Koch, *Bayes Theorem*. Springer, 1990.

[24] D. L. Faigman and A. Baglioni Jr, "Bayes' theorem in the trial process: Instructing jurors on the value of statistical evidence." *Law and Human Behavior*, vol. 12, no. 1, p. 1, 1988.

[25] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[26] B. Kosko, "Neural networks and fuzzy systems: a dynamical systems approach to machine intelligence/book and disk," *Vol. 1Prentice hall*, 1992.

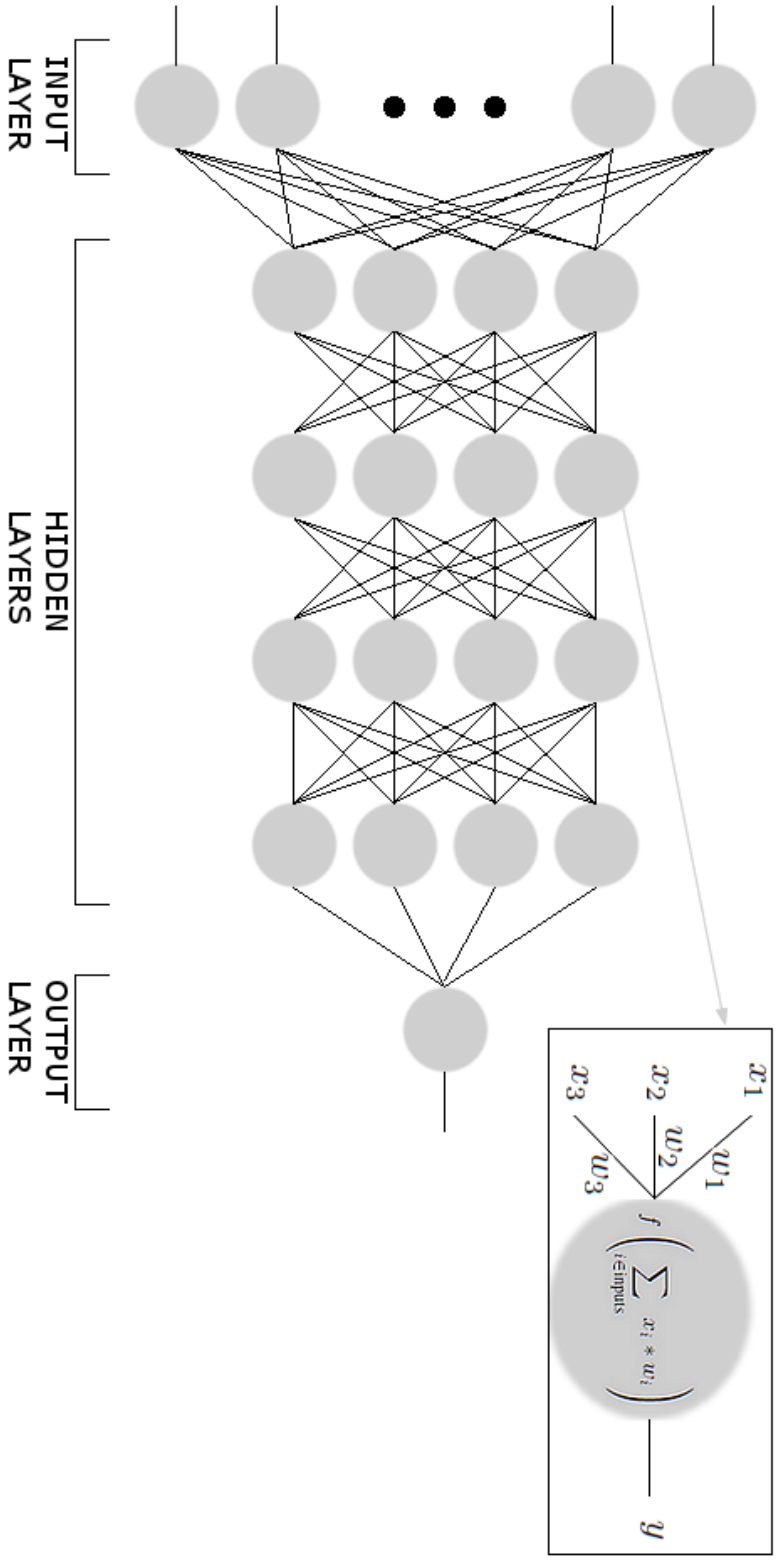

Fig. 6: The model of the proposed artificial neural network with 4 hidden layers.

[27] D. F. Specht, "Probabilistic neural networks," *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.

[28] P. d. B. Harrington, "Sigmoid transfer functions in backpropagation neural networks," *Analytical Chemistry*, vol. 65, no. 15, pp. 2167–2168, 1993.

[29] H. Yonaba, F. Anctil, and V. Fortin, "Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting," *Journal of Hydrologic Engineering*, vol. 15, no. 4, pp. 275–283, 2010.

[30] M. Panicker and C. Babu, "Efficient fpga implementation of sigmoid and bipolar sigmoid activation functions for multilayer perceptrons," *IOSR Journal of Engineering (IOSRJEN)*, pp. 1352–1356, 2012.

[31] R. Hecht-Nielsen, "Theory of the backpropagation neural network," in *Neural Networks, 1989. IJCNN., International Joint Conference on*. IEEE, 1989, pp. 593–605.

[32] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The rprop algorithm," in *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993, pp. 586–591.

[33] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.