KNOWLEDGE – International Journal
Vol. 23.5
Budva, Montenegro, May, 2018

STUDY ON THE PROCESSING OF NATURAL LANGUAGES (ALBANIAN LANGUAGE) THROUGH METHODS OF ARTIFICIAL INTELLIGENCE

Slavço Çungurski
Faculty of Informatics, FON University, Skopje Macedonia, slavcho.chungurski@fon.edu.mk ,
Burim Rexhepi
Faculty of Informatics, FON University, Skopje Macedonia, burimrexhepi007@gmail.com,
Dimche Zvrcinov
Faculty of Informatics, FON University, Skopje Macedonia, dimce012@gmail.com

Abstract: It has been years now, since the most developed countries in the world and the most spoken languages, included the Natural Languge Programming discipline in the computer science research process. This discipline, in difference with many others, even though includes general principles serving all languages of the world, has a difficulty to standardize their practical usage for all the natural languages. This happens partially because of the specifics that different languages have (including morphological, syntactical, grammatical and semantical ones), partially because of the dynamics (such as the constant language evolution), but also
Keywords: Natural language processing, speech recognition, neural network SOM, FST, Albanian language, training, MFCC, HMM.

1.INTRODUCTION TO THE STUDY

Natural language processing includes methods, techniques and tools for processing natural languages. Natural languages include two basic groups: spoken and spoken languages written. Both of these groups are intertwined with one another. But there are also somedifferentiations. The differences relate to the provincial, ethno - cultural and personal characteristics that there is the use of spoken language. Often times in the spoken language dictionary are used barbarism, parasitic parts of the speech, words expressed differently as a result ofused dialect, words unsealed in literary standard but used in the provincescertain, professional expressions etc. We should not forget the provincial features offont expression in different areas (eg: - a nose, nth, and xh or ç and q inthe geg dialect, the addition of the letter between the two concurrent consonants in the dialect tosk as:film instead of film etc.) or individual characteristics related to featuresverbal forms (people who speak r as in the mouth, talk peoplewhistling letters as a result of dental crown deformation etc.).Additionally, natural languages can not be introduced through machines with a conditionas is the case with programming languages. Natural languages are combined in waysvarying by presenting an infinity of combinations of different syntax forms.On one hand the richness of the vocabulary in the natural languages captures several thousand to several dozen thousands of words, exceeding several times the number of words or phrases of a language programming and in turn a huge number of their positioning combinationsexpress the same semantics, or repositioning the sam word in contexts different to give different meanings is enough to express the difficulty of great that has the enterprise of informatics of a natural language. Exactly for natural language features have begun to be processed relatively late and are included in the Artificial Intelligence Informatics discipline. Last but not least is the specificity of each of the languages natural has. Like the programming languages, the number of natural languages is also extremely large. "It is estimated that the number of languages spoken around the world ranges up to 7,000. 90% of these languages are used by less than 100,000 people. According to these data we can say that languages that actually have the opportunity initially develop their computer processing can be in the 600th order languages. This large number of languages indicates that, despite the features of common methods and techniques they can use, a large work volume it remains to develop individually for each of these natural languages. This task too intense that is related to each of the natural languages has been one of the obstacles greater in the processing of many natural languages, including language albanian.

2. NAMES IN ALBANIAN

"The name is called that part of the vocabulary that names living things and things and does grammatical category such as gender, number, race, and excellence.
This is the standard definition of the name, so what the name changes is syntactic compatibility with the different parts of the sentence in gender, number, rank or distinction. The Albanian language has three races: masculine, feminine, and neutral, where the latter is almost unusable. Gender is one of the elements that appears in both numbers, as well consequence is one of the key elements to be identified in one language. But sex

does not make it difficult to identify the name as it remains the same in all numbers, rasat or forms. The number is an element that generally affects all of the general names. The sum number can be obtained by changing the default or changing the subject itself of the name. (eg bird - birds, old - age, night - net).

3.KNOWLEDGE OF ALBANIAN SPEAKING LANGUAGE METHOD THROUGH THE MARKOV MODEL OF THE SECRET

The Method of Recognition of Speech Language by the Markov Model (HMM) is the most commonly used way to Speech Language Recognition. The Origin of this method is related to the recognition of the force of earthquakes or explosions starting from
the echoes of the waves they left behind [6]. However its wider use today is on
the classification of the sound waves produced during the speech process. This process is known as Lecture Recognition or Knowledge of Speaking Language is one of the most important tasks today difficult to find information technology. Desire and need for simultaneous writing of spoken text relates to a psychological process of man himself which is the passage through writing i its existence. Below is an introduction to the Markov Secret Model .

4.MULTILAYER PERCEPTRON

The first and oldest model built on neural networks is what is called Multilayer Perceptron. This network model is a finite three-layered acyclic graph defined as: Input, Output Layer, and Hidden Layer. Hidden layer is a theoretical layer because actually the number of layers hidden under this layer may be more than one. There is even a pattern of perceptron with many supplements where the middle layer is completely missing. The underlying principle of this neural network model is that the i-i add-on output serves as an input for the (i + 1) add-on. For this reason, the model is used for numerous tasks related to various disciplines of artificial intelligence, including the Recognition of Lectures.
The figure below represents a conceptual graphical model of multi-layered Perceptron.
The following Multiple Layer Perceptron is referring to the MLP acronym

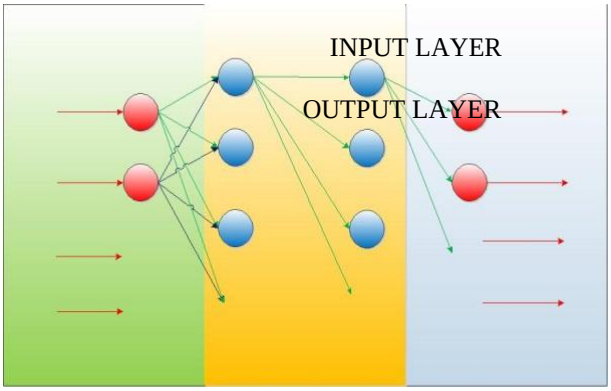

Figure 1. Multiple Layer Perceptron is referring to the MLP acronym

5.RECURRENT NEURAL NETWORKS (RNN)

A repeating network is a neural network with a retrograde connection. According to Mike Shuster there are several different types of Neural Repetitive Networks that differ specifically in input data such as: Simple RNN, Single Signature RNN (URNN) which is composed of two subgroups: Forward RNN Delayed and Backward RNN Delayed, as well as the RNN Guidelines (BRNN). As can be seen all these groups that in the name represent different from each other the training algorithm used as well as whether the input data contain examples from the examples posted during the study of the signal. RNNs are an alternative solution based on supervised training in Pattern Recognition. Also in Lecture and Voice Recognition has better results and an increase in recognition rate than MLP. The main idea behind the RNN is to not only use the actual frame but also the frames previously processed (and sometimes the frames that are expected to come next) to find access to a frame with a particular letter. One of the main activities performed by the system on a neural network to recognize a specific motive is the training method and the applicability of this training. As we said above, RNN training is supervised. This training method similar to that used in an MLP includes:
• collecting data samples from different sources - different people and materials

• Identification between data (as it is specifically spoken about Lecture Recognition - waves from recording the lecture) which frame is approaching which letter
• initiate network training under the predefined algorithm of specific architecture using the randomly selected RNN mode.

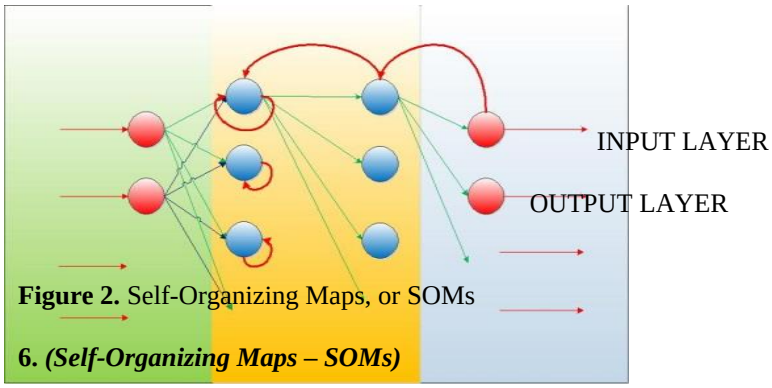

Figure 2. Self-Organizing Maps, or SOMs

6. *(Self-Organizing Maps – SOMs)*

Self-Organizing Maps, or SOMs, are a category of neural networks that are highly used in the process of recognizing the motive. This algorithm has a particularity compared to the two neural networks mentioned above. His
learning process is done without supervision.

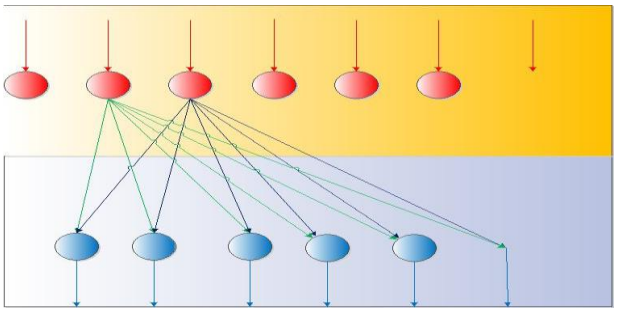

Figure 3. Map of Self-Organizing Maps.

7.CONCEPTUAL MAP OF SELF-ORGANIZING MAPS. SOMS WORK DIFFERENTLY FROM THE FEEDFORWARD NEURAL NETWORKS

Generally a neural network includes a layer hidden in the sketch of its architecture, SOMs do not have this layer that makes them lighter than other networks. An SOM also changes at the entrance and exit from other neural networks. In terms of entry, SOMs are fed with information pieces derived from a data collection process, which in the case of Lecture Recognition would be the production of the Vocal Human Devices. Since the generated signal is continuous, it must first be quantized. Then the quantized result will be processed by obtaining some characteristic coefficients where most known of which are called MFCC and LFCC. In the case of Lecture Recognition, depending on the method used to process the message, as mentioned before, entry may have a different number of neurons in support of incoming information. When producing MFCC coefficients, the number of coefficients produced for each time unit is generally 12, 13 or 39. Depending on this number, the number of neurons in the entry layer will be 12, 13 or 39 respectively. In the case of this dissertation, to reduce the complexity of the input information, the number of neurons at the entrance will be 12. In the output layer the system should have as many neurons as is the number of possible different results that can be obtained from the combination of information in the incoming neurons. Since the Albanian language has 36 letters and even phonemes, then 36 should be the number of outgoing neurons that would be used by the SOM neural network
to solve the problem of Lecture Recognition. If we referred to Jurafsky D. and Martin J. [D5] "The most common configuration is to use 3 HMM states: an input, a middle, and an output. Each of the sounds consequently consists of three HMM broadcasting states and not one single. " It would be simpler to use the same approach to selecting the number of neurons in the outgoing layer of a SOM network. This would make it easier to orient the different parts of a phoneme sampled with the sub-phoneme on the right node and not run a

subframe to a wrong output path. At the end of the system, the neurons sharing the same phoneme can be reunited into a single letter count. This model could be simplified by uniquely designating vowels as they have a harmonic signal. The vowel feature can also handle a sampling at a shorter time interval than the consonants [15]. But this approach also presents its own difficulties. First, is it proven experimentally that a subfound state system is more efficient than a system that equates to a phoneme output node. We will talk about it later. And secondly, increasing the number of neurons in the exit layer can greatly increase the difficulty in system training as the number of outbound nodes directly affects cardinality of the problem. Since the number of outbound joints is large, a large number of characters (thousands or tens of thousands) need to be trained to fully train the system. Sphinx - 4 suggests to be trained by a single person's voice for at least 1 hour to enable voice commands and controls to be recognized and 5 hours material recorded by 200 speakers in case we want to have the full knowledge of the Lecture by an anonymous speaker . SOMs based on the data used for training will activate in the output a neuron that has the best alignment with the motives identified by the input data. As mentioned earlier, the algorithm used to train SOM networks is unattended, which means that there is no need for human support during the training process. The learning process is associated with a learning report that is specific to neural networks, including SOMs, and ranging in interval $0 < k < 1$.This value is solved empirically and is mainly among the numbers $0.4 < k < 0.5$. Then this value is used in the weighing process (which can be added or removed) and certain values can improve the process of recognizing a motif. SOM has long been being used successfully in handwriting recognition [21], [22]. The concept of Knowing the Lecture should not be too much of a change. Suffice it to replace the motives associated with the local light intensity differences

8.CONCLUSIONS

Look at the search

As a conclusion of this dissertation, it is very important to mention the concrete contribution given in the field of information technology. Since the processing of native language in Albanian is in its early stages, contributions also include an important part of the theoretical and practical foundation for achieving a fuller elaboration of standard written or spoken Albanian languages. Part of the concrete contributions to this paper is the general digitalization of the Albanian language vocabulary which, although not a fully-fledged vocabulary, can cover a good part of the words that the Albanian language includes.

I think that since every problem of natural language processing begins with the examination of a digital dictionary, it is the task of Albanian language institutions to devise such a dictionary and make it accessible to the public free of charge. Such work has already been done for all the important languages of the world and is a necessity for the Albanian language. Such a certified vocabulary eases the work of researchers and leaves no case for equivocation. Moreover, products supported on a purged dictionary can also be used as commercial products. Another important contribution may be the construction of an FST for the recognition of translated forms of Albanian language. Certainly, this FST can be further elaborated and further enriched by the exception model by adding irregularly developed words (plural of irregular or spiritual names, partial or plain verbs, auxiliary verbs, etc.).). This task is manual and requires a great deal of discipline in building the exception file. These exceptions can be used for different tasks and may also be included in the dictionary as special forms to keep the smallest file of special rules.

The model used for FST construction is completely unique and unprecedented in literature. So it's not just an application of a FST-ready model applied to Albanian, but it is a model built around a language, which is in this case the standard Albanian, and can be used for other languages too which have a considerable amount of deigned forms and special cases. As an example of languages that can use this algorithm, they can be considered different languages from English, which encompasses all the special word terms in the dictionary and can use an extremely simple algorithm to identify words . Languages that can use this algorithm are languages like those of Latin origin - in the case of verbs, Slavic - in the case of verbs and names, Greek as well - in the case of verbs and names.

9.ACHIEVEMENTS

Of course, the main contribution this is building a model and its practical application in the recognition of the Albanian spoken language. This task, often essential in the literature related to the processing of natural languages, is one of the most difficult tasks not only theoretically but also in its practical realization. On one hand, the recognition of numbers in Albanian language (including large numbers) was realized, and a lexical analyzer for the numbers in Albanian was also built. On the other hand, it is very interesting in the scientific field to construct it using neural networks of the SOM type, which are not only included in the group of advanced methods in the theory of machine learning but also studied mainly by theoretical (at least no genuine practical language recognition applications were found) for Lecture Recognition training.Also within this

system, a model was proposed that helps strengthen the system, based on the recognition of the voice signal in letters / syllables rather than simply in letters. This system, increasing the amount of information, reduces system's indeterminacy, thus making the system more robust. On the other hand, it was found that this system is very difficult to build in the absence of human resources. Since the analysis of the neural network outputs at the output for the letters is a proving and continuous improvement, and given that the number of letters is only 36 and that of the syllables is several hundred, let's think of the little time it takes for identifying all combinations of exits from neural networks for these hundreds of syllables. The system built during the training period and the building of the grip - exit took about a month, let's count the time it takes to analyze and record these hundreds of syllables (24 grips were taken for the numbers).Full language recognition systems are scientific obligations of scientific institutes and research centers or groups. The work of this doctorate is best not to be forgotten but to continue to expand, enrich and improve. Future researchers may initiate the problem of Lecture Recognition and either provide alternative solutions to compare or exceed this system or improve this system to a better result. Such a complete system would have scientific but also commercial benefits if it were to be included as a module in larger systems. Even a system that only recognizes numbers can be used in commercial products (eg an IVR that also receives voice commands, a sound calculator, etc.). One can further think of the human or economic benefit of a system that knows all the

10.FUTURE WORK ON THIS TOPIC

In addition to the developments developed for the Lecture Processing process, another important work is the processing of Lecture Synthesis, that is, the sound rebuilding of a written text or what is commonly called teskst lecture. This process, seemingly simpler than Lecture Recognition, actually involves major technical difficulties in achieving fluidity and excellence in voice production. Even more often, literature refers to a problem that can be solved after Recognition of the Speech as the Speaker can automatically do that part that otherwise needs to be done manually that is, the division of the voice signal into the parts that are recorded and then used for voice creation.At the conclusion of the whole dissertation we can conclude that advancing and resolving the problems of the Processing of Natural Languages is a great step towards the automation of human affairs. Since man exchanges with the environment that surrounds about 25% of the information between speech and listening, the basic tool of which is natural language, the computerization of this process brings more man to the car and makes the latter more family and warm for her usewords of the Albanian language. Of course, that so complicated systems are still far away. Even the largest languages in the world are far from the problem of knowing more than one word in a row. Research that is being made today around the world is oriented towards this field. This is evidenced by the companies that are running to get the exclusivity of an original method of patenting it later. Even in this prism, university institutions should feel more motivated in the race, leaving such a discovery in the hands of citizens who have this great knowledge of mankind

REFERENCES

[1] Languages of the world. http://www.bbc.co.uk/languages/guide/languages.shtml
[2] Indo – European Languages – Evolution and Locale maps. Slocum J., Linguist Research Center, The University of Texas at Austin & Collage of Liberal Arts, http://www.utexas.edu/cola/centers/lrc/general/IE.html
[3] The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking. Bogert B.P., Healy M.J.R., dhe Tukey J.W., Time Series Analysis, M. Rosenblatt, Ed., 1963, kap.15, f. 209–243
[4] Statistical Methods for Speech Recognition. Jelinek F., Bradford Book, 1998, f.1 -305
[5] Markov Models for Pattern Recognition. Fink G. , Springer, 2003, f. 1 - 248
[6] Speech and Language Processing, 2nd Edition. Jurafsky D., Martin J., Prentice Hall, 2009, f.1 - 1024
[7] Sphinx-4: A Flexible Open Source Framework for Speech Recognition – Whitepaper. Walker W., et al., 2004