Computer Science Journal of Moldova, vol.18, no.1(52), 2010

Determination of inflexional group using P systems

Svetlana Cojocaru, Elena Boian

Abstract

The aim of this article is to describe the process of determining the inflectional group using P systems with replications. In this process firstly the sets of endings of the same length are constructed, to which inflectional models are put into correspondence. Based on these endings the inflectional model for arbitrary word is determined.

Introduction

Natural Language Processing (NLP) is one of the areas that requires high performance computing. In order to solve problems in this domain, it needs to operate with resources containing millions of entries, so there is a logical temptation to apply approaches based on parallelism. Such attempts have been undertaken since the '80s, and have an ample development in coming decades [1, 2].

An important direction in NLP is creation of computational linguistic resources. In [3] we presented the solution of one of the problems that contributes to resources enrichment: automatic inflection. The proposed solution was applied for the Romanian language. Thanks to this process, we received over 40 word forms for each verb, 24 new words for each adjective, etc. Our solution was based on the use of P systems with replication and applied for the case when inflectional model is known a priori. As a rule, there are classification dictionaries for high inflectional languages, where these models are established. In the case of the Romanian language we use the dictionary [4], containing the inflectional models defined for about 30,000 words. In this paper it is shown how the inflection model for an arbitrary word can be determined. Knowing this information we are able to perform automatically the inflectional process. Analogously to [3, 5] for illustration we will use examples from the Romanian language, but the proposed method can be applied also to other natural languages with similar inflectional mechanisms.

From the beginning we must note that in general case an algorithmic solution to this problem is not possible. The first obstacle is the determining of part of speech: there are a lot of examples of homonyms which denote different parts of speech (e.g.: *abate* – masculine noun (engl. abbot) and verb (engl. to divert).

Let us restrict the formulation of the problem: is it possible to ascertain inflectional model knowing the part of speech? The answer is negative in this case too.

For confirmation there is a list of examples which prove that we can not determine the inflectional model without invoking the etymological or phonetic information.

This assertion can be illustrated by analysing feminine noun *masă*. Following the meaning of the furniture object we form the plural *mese* (engl. *tables*) using the inflectional model with vowel alternation "$a \rightarrow e$". But if the meaning is *mass* [6], plural *mase* will be produced without vowel alternation. The origin of this phenomenon is etymological: the first case is of the Latin origin *mensa*, and the second – the French word *masse*) [6].

But the problem can be tackled in other mode: we can establish certain criteria that allow us to conclude in the term of word structure analysis, if it is possible to determine the inflectional model or not, and in the case of "yes", to determine which is namely the respective model. Otherwise speaking, we try to formulate the criterion under which we can say that inflectional process is performed automatically and can indicate the appropriate inflectional model.

1 Problem of inflectional model determination for an arbitrary word

The specific character of the investigated area (natural language) is reflected in the fact that many of the objects and concepts it operates, cannot be the subject to strict formalisation. Therefore we will try to distinguish certain classes in which this formalisation is possible.

So let the word-lemma be known in its graphical representation (i.e., the data without phonetic, etymological notes, etc.). Also let the part of speech be known, and for nouns – the gender. We divide the words into three categories: irregular, absolutely regular and partially regular.

For all parts of speech the fact of belonging to the *irregular* class is determined by the fact of their belonging to a set of words known a priori. To simplify the statement we exclude from the examination the set of irregular words, their presence (or absence) does not affect the generality of the algorithm. We consider *absolutely regular* words, to which a single inflectional model corresponds and we note by A the set of their endings. We call *partially regular* those words, to which two or more inflectional models correspond. The set of their endings we denote by P. In the following we establish the criteria for belonging to these two classes (and corresponding inflectional models). The algorithm described in [7] determines these criteria in sequential mode. We propose to obtain these criteria in parallel mode using massive parallelism which is characteristic to membrane P systems [8].

Inflectional group determination will be made in two steps:

- building the sets of endings of the same length, to which the inflectional models are being put into correspondence;

- determination of the inflectional group in correspondence with the built sets of endings.

2 Construction of sets of endings

Let L be the set of all words of a language. We come from the assumption (valid for majority of natural languages) that there is a classification dictionary $D \subseteq L$, so that to any $\omega \in D$ it puts into correspondence an inflectional model ν, where ν is a positive integer. We will present dictionary D as a union of words classified by parts of speech (and gender, for nouns), $D = \cup(C)_{i=1}^{5}$, where C is one of the sets of words, which belong to the open classes [3] (for Romanian these are the adjectives, verbs, nouns: masculine, feminine, neuter). For each C_i the dictionary D puts into correspondence the finite set of inflectional models $N_i = \{\nu_1, \ldots, \nu_{n_k}\}$, such that for $\forall \omega \in C_i$ there is at least a $\nu \in N_i$. We will separately operate with each of these classes.

Let C be one of these classes. The idea of algorithm to build the sets of endings is the following. For each word $\omega \in C$, to which the inflectional model $\nu_m \in N$ corresponds (N is the set of integers of inflectional models for words in C), there are built the endings with decreasing lengths from $|\omega|$ to 1. The pairs (γ_i, ν_m) are formed, where γ_i is a substring of length i of the word ω, $(1 \leq i \leq |\omega|)$. The pairs, constructed thus, are compared and filtered. The filtration process is carried out in the following way: out of each two elements (γ_i, ν_m), (η_i, ν_n), we keep only one, if $\gamma_i = \eta_i$ and $\nu_m = \nu_n$, where γ_i is a substring of length i of the word $|\omega|$, and η_i is a substring of length i of the word ψ (i.e. only noncoincident pairs are kept).

If for all the pairs in which $\gamma_i \neq \eta_i$ the equality $\nu_m = \nu_n$ takes place, then the pairs (γ_i, ν_m) and (η_i, ν_n) are elements of the set A of the endings corresponding to absolutely regular words.

If $\gamma_i = \eta_i$ and $\nu_m \neq \nu_n$, then the ending η_i indicates a substring of the word ψ partially regular from the set P, to which several inflectional models $\nu_m, \nu_n, \ldots$ correspond.

We denote by $L_{max} = max\{|\omega|\}$, $\omega \in C$, maximum of the length of words in C.

This algorithm can be realised using the following membrane system Π_1.

$$\Pi_1 = (O, \Sigma, \mu, R_0, R_i, A, P),$$

where $i = 1, ..., L_{max}$,

O is the alphabet of symbols, λ is an empty element, $\lambda \in O$.

$\Sigma \subseteq O$ – Romanian alphabet,

μ – membrane structure which is defined as:

$$\mu = [_0\, [_A\,]_A\ [_{L_{max}}\ [_{L_{max-1}} \cdots\ [_1\ ]_1\ \cdots]_{L_{max-1}}\ ]_{L_{max}}\ \ [_P\ ]_P\ \ ]_0,$$

R_0: $\{(\omega, \nu_m) \to (\omega, \nu_m)^1 || \ldots || (\omega, \nu_m)^i || \ldots || (\omega, \nu_m)^{|\omega|}$, $1 \leq i \leq |\omega|$, $\omega \in C$, $|\omega| \leq L_{max}$,

$(\omega, \nu_m)^i \to ((\omega, \nu_m)^i, in_i)$, for $i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}\}$;

R_i: $\{(\omega, \nu_m)^i \to (\gamma_i, \nu_m)$, for $i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}$, where $\omega = \omega_l \gamma_i$, $|\gamma_i| = i$, $i = 1 \ldots |\omega|$,

$(\gamma_i, \nu_m) \to \lambda\ |_{\exists(\eta_i, \nu_n): \gamma_i = \eta_i \& \nu_m = \nu_n}$, $i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}$, $\psi = \psi_l \eta_i$, $|\eta_i| = i$, $i = 1 \ldots, |\psi|$, $|\psi| \leq L_{max}$, $\omega, \psi \in C$,

$(\gamma_i, \nu_m) \to (\gamma_i, \nu_m, \nu_n, \ldots,$
$\ldots, \nu_{n_k})\ |_{\exists(\eta_i^1, \nu_{n_1}), (\eta_i^2, \nu_{n_2}), \ldots, (\eta_i^{n_k}, \nu_{n_k}): \gamma_i = \eta_i^1 = \eta_i^2 = \ldots = \eta_i^{n_k} \& \nu_m \neq \nu_n}$,
$i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}$, $i = 1 \ldots, |\psi|$, $|\psi| \leq L_{max}$, for $\forall k = 1, \ldots, j$,

$(\gamma_i, \nu_m) \to ((\gamma_i, \nu_m), out_A)\ |_{\forall \eta_i: \eta_i \neq \gamma_i}$, $i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}$, $i = 1 \ldots, |\psi|$, $|\psi| \leq L_{max}$,

$(\gamma_i, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k}) \to$
$((\gamma_i, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k}), out_P)\ |_{\exists \eta_i^1, \eta_i^2, \ldots, \eta_i^{n_k}: \eta_i^1 \neq \gamma_i \& \eta_i^2 \neq \gamma_i \ldots \& \eta_i^{n_k} \neq \gamma_i}$,
$i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}$, $i = 1 \ldots, |\psi|$, $|\psi| \leq L_{max}$, for $\forall k = 1, \ldots, j\}$.

The membrane A contains objects of type (γ_i, ν_m).

The membrane P contains objects of type $(\gamma_i, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k})$.

The rule R_0 indicates the replication of objects $(\omega, \nu_m)^i$ for $|\omega|$ times. Each object $(\omega, \nu_m)^i$ is transferred into the region bounded by the membrane i ($i = 1, \ldots |\omega|$).

The rule R_i indicates the following:

– truncation of the word ω keeping the ending γ_i of the length i,

– elimination of the pair (γ_i, ν_m) from the region i in case if there exists the duplicate pair. When such duplicate pair does not exist, the remained object is transferred to the membrane A,

– in case if the same ending of length i has in the capacity of pair different numbers of the inflectional models, then the object of the type $(\gamma_i, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k})$ is formed, which is transferred into the membrane P.

Finally in the membrane 0 two resulting membranes A and P are obtained containing criteria for setting inflectional models for absolutely regular and partially regular words.

3 Determination of the inflectional group

We will determine the inflexional group for the word $\psi \in C$.

The idea of algorithm for the inflexional group determination is the following.

The substrings ξ_i $(1 \leq i \leq |\psi|)$ of the endings with decreasing length from $|\psi|$ to 1 of the word ψ are constructed. Initially we look for a completely regular model, comparing the ending ξ_i $(|\xi_i| = i)$ with the elements $(\gamma, \nu_m) \in A$ $(|\gamma_i| = i)$. If $\exists \gamma_i = \xi_i$, then ν_m is the inflectional model number. In case if we did not find an appropriate model in A, we look for it in P. If $\exists \gamma_i = \xi_i$ $(\gamma_i, \nu_{n_1}, \nu_{n_2}, \ldots, \nu_{n_k} \in P)$, the word ψ is partially regular and it has to inflect in correspondence with the inflexional models $\nu_{n_1}, \nu_{n_2}, \ldots, \nu_{n_k}$. In the case when $\xi_i \neq \gamma_i$ for $\forall \gamma_i$ from A and P the inflectional model can not be determined automatically and the intervention of user (the expert in linguistics) is needed.

This algorithm will be described by the following membrane system Π_2.

$$\Pi_2 = (O, \Sigma, \mu, A, P, R_0, R_i, 0),$$

where $i = 1, \ldots, L_{max}$,

O is the alphabet of symbols, including element "*false*"$\in \Sigma$,

$\Sigma \subseteq O$ – is the Romanian alphabet. The element "*false*" will signal about the case when the inflectional model is not found for the word to be inflected.

μ – membrane structure which is defined as the following:

$$\mu = [_0\ \ [_A]_A\ \ [_{L_{max}}\ \ [_{L_{max-1}}\ \cdots\ \ [_1\ \ ]_1\ \ \cdots\ \ ]_{L_{max-1}}\ \ ]_{L_{max}}\ \ [_P]_P\ \ ]_0,$$

A contains the set of pairs of type (γ_i, ν_m) own to absolutely regular words ($i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}, k = 1, \ldots, |C|$).

P contains the set of elements of type $(\gamma_i, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k})$ belonging to partially regular words ($i = 1, \ldots, |\omega|$, $|\omega| \leq L_{max}, k = 1, \ldots, |C|$).

R_0: $\{\ \psi \rightarrow (\psi, in_1)||\ldots||(\psi, in_i)||\ldots||(\psi, in_{|\psi|})$,

$\psi \rightarrow (\xi_1, in_1)||\ldots||(\xi_i, in_i)||\ldots||(\xi_{|\psi|}, in_{|\psi|})$, where $\psi = \omega_j\xi_i$, $|\xi_i| = i$, $i = 1, \ldots, |\psi|\ \}$,

R_i: $\{\xi_i \rightarrow ((\psi, \nu_m), out_0)\ |_{\exists(\gamma_i,\nu_m)\in A:\xi_i=\gamma_i}$, $i = 1, \ldots |\psi|$, $i = 1, \ldots, |\omega|$, $|\psi| \leq L_{max}$, $|\omega| \leq L_{max}$,

$\xi_i \rightarrow ((\psi, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k}), out_0)\ |_{\exists(\gamma_i,\nu_m,\nu_n,\ldots,\nu_{n_k})\in P:\xi_i=\gamma_i}$, $i = 1, \ldots |\psi|$, $i = 1, \ldots, |\omega|$, $|\psi|$, $|\omega| \leq L_{max}, k = 1, \ldots, |C|$,

$\xi_i \rightarrow ((false), out_0)\ |_{\xi_i\neq\gamma_i}$, for $\forall(\gamma_i, \nu_m) \in A \vee (\gamma_i, \nu_{n_1}, \nu_{n_2}, \ldots, \nu_{n_k}$ $\in P$, $i = 1, \ldots |\psi|, k = 1, \ldots, |C|\}$.

The rule R_0 indicates replication for $|\psi|$ times of the endings ξ_i ($\psi = \omega_j\xi_i, |\xi_i| = i$) which are transferred then to membrane i ($1 \leq i \leq |\psi|$).

The rule R_i forms objects of type $(\psi, \nu_m) \in A$ or $(\psi, \nu_m, \nu_{n_1}, \ldots, \nu_{n_k})$ $\in P$ by which one or more inflectional models are assigned to the word ψ. The formed objects are transferred to the external membrane.

The appearance of the value "*false*" in the region 0 means that the number of the inflectional model is not found for the word $\psi \in C$. In this case the inflectional model can not be determined automatically.

4 Example of using membrane systems Π_1 and Π_2

Let D = { *(grup,1),(grup,2), (dulap,1), (cuvânt,2), (vânt,1), (tractor,3), (muzeu,41)*}.

Initially $A = \emptyset$, $P = \emptyset$ (see Fig.1).

We will take as C all the words from D, i.e.,

C = { *grup, dulap, cuvânt, vânt, tractor, muzeu*}

(in English: group, wardrobe, word, wind, tractor, museum).

$L_{max} = 7$; $N = \{1, 2, 3, 41\}$.

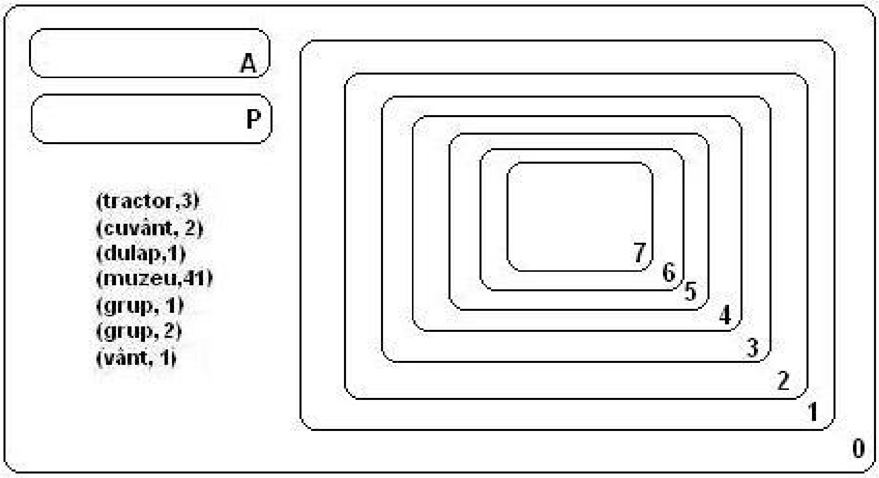

Figure 1. Initial state of membrane system Π_1.

The Figure 2 illustrates the process of building the sets of endings of the same length of words from C, to which the inflectional models N are being put into correspondence.

We obtained the sets A and P with the following components (see Fig.3):

A = { *(dulap,1), (ulap,1), (lap,1), (ap,1), (cuvânt,2), (uvânt,2), (tractor,3), (ractor,3), (actor,3), (ctor,3), (tor,3), (or,3), (r,3), (muzeu,41), (uzeu,41), (zeu,41), (eu,41), (u,41)* }.

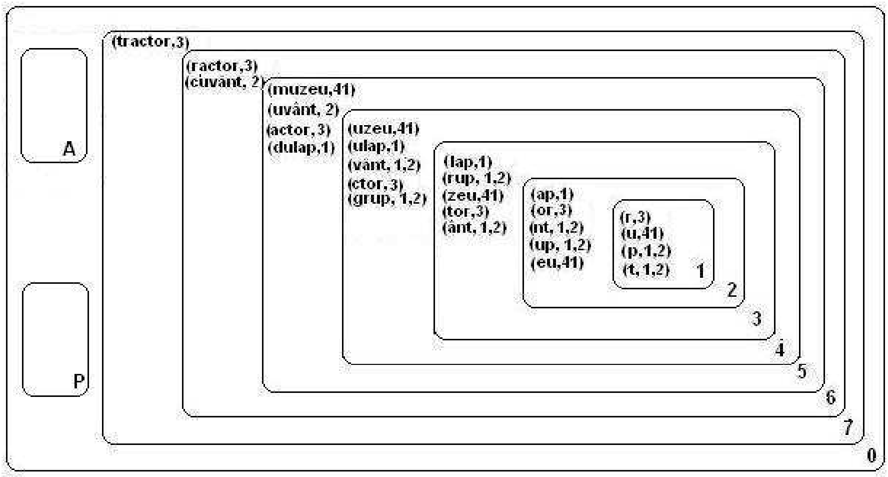

Figure 2. Building the sets of endings of the same length for words from C.

$P = \{$ *(grup,1,2), (rup,1,2), (up,1,2), (vânt,1,2), (ânt,1,2), (nt,1,2), (p,1,2), (t,1,2)* $\}$.

The Figure 3 illustrates the process of determining the inflectional group for the word *motor* (in English: engine) in relation to the built sets of endings.

We obtained that the word *motor* will be inflected using the inflectional model 3 (Fig. 4).

Conclusions

On the basis of the classification dictionary D for each part of speech C the sets A and P are constructed, which determine the inflexional models for absolutely and partially regular words. We mention the following.

1. In the general case we can renounce to build the respective sets for each part of speech apart, but in this way the number of partially

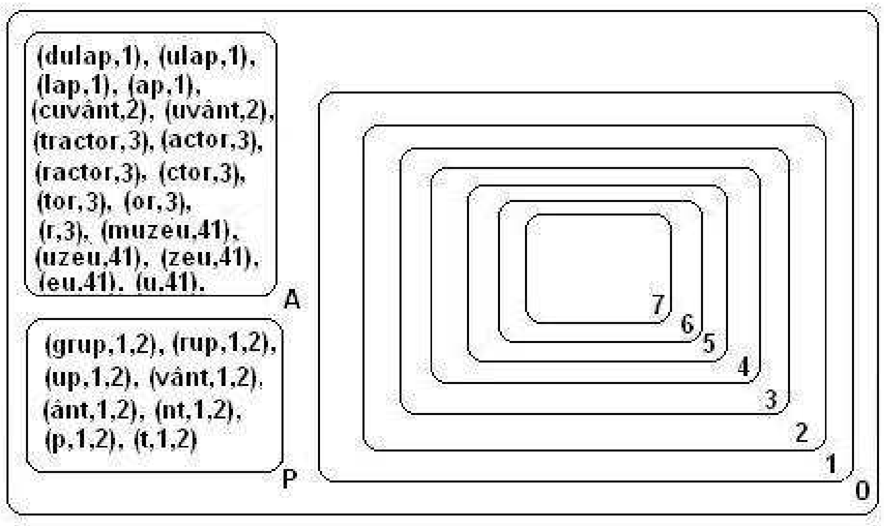

Figure 3. Obtaining the sets A and P.

regular words would be increased.

2. Both systems Π_1 and Π_2 could be reduced only to the membrane 0 with two membranes A and P contained in it. The inner membranes i were introduced in order to separate the comparing processes, which would allow us to simplify implementation of such mechanisms by a simulator.

3. The experiments, performed for a set of about 2000 base words, showed that in 97% of cases the inflectional model number can be determined using the systems described above. The 3%, for which the result was marked by "*false*", present in most cases the irregular words.

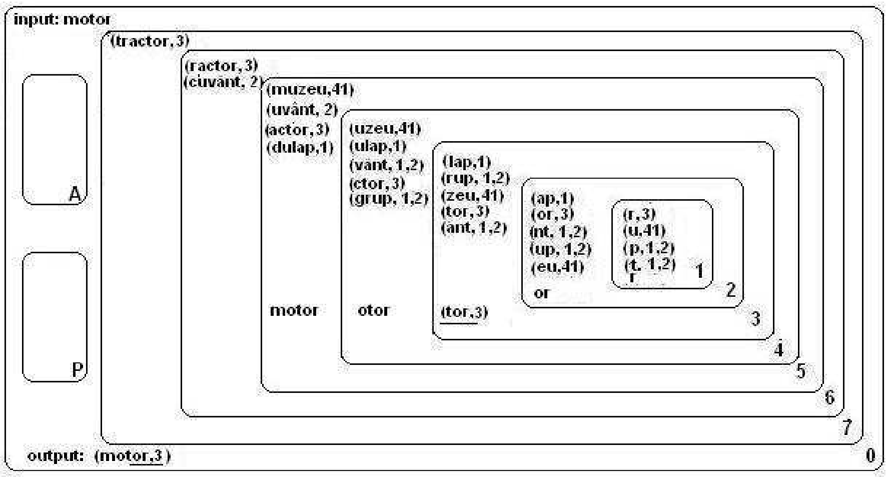

Figure 4. Example of using the membrane system Π_2

References

[1] Eiichiro Sumita, Kozo Oi, Osamu Furuse, Hitoshi Iida, Tetsuya Higuchi, Naotao Takahashi, Hiroaki Kitano. *Example-Based Machine Translation on Massively Parallel Processors.* Proceedings of the 13th International Joint Conference on Artificial Intelligence, Chambery, France, 1993, vol.2, pp.1283–1289.

[2] Hiroaki Kitano. *Challenges of Massive Parallelism.* Proceedings of the 13th International Joint Conference on Artificial Intelligence, Chambery, France, 1993, vol.1, pp.813–834.

[3] A.Alhazov, E. Boian, S. Cojocaru, Yi.Rogojin. *Modelling Inflexions in Romanian language by P Systems with String replications.* Computer Science Jourmal of Moldova, v.17, 2(50), 2009, pp.160–178.

[4] A.Lombard, C.Gâdei. *Dictionnaire morphologique de la langue roumaine.* Bucureşti, Editura Academiei, 1981, 232 p.

[5] S.Cojocaru. *Romanian lexicon: Tools, implementation, usage.* In Tufis, D., Andersen, P., eds.: Recent Advances in Romanian Language Technology. Volume I., Editura Academiei (1997), pp. 107–114 ISBN 973-027-00626-00.

[6] www.dexonline.ro

[7] S.Cojocaru. *The ascertainment of the inflexion models for Romanian.* Computer Science Journal of Moldova, vol.14, N1 (40), pp.103–112.

[8] Gh. Păun, *Membrane Computing. An Introduction*, Natural computing Series. ed. G. Rozenberg, Th. Back, A.E. Eiben, J.N. Kok, H.P. Spaink, Leiden Center for Natural Computing, Springer - Verlag Berlin Heidelberg New York, 2002, 420 p.

Svetlana Cojocaru, Elena Boian,

Received June 8, 2010

Institute of Mathematics and Computer Science,
Academy of Sciences of Moldova
Academiei 5, Chişinău MD-2028 Moldova
E–mail: {Svetlana.Cojocaru,lena}@math.md