საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 7, №1, 2013
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 7, no. 1, 2013

Linguistics

The Net Representation of Interactive Language Processor

Giorgi Chikoidze*, Ana Chutkerashvili*, Nino Javashvili*

* *Archil Eliashvili Institute of Control Systems of the Georgian Technical University, Tbilisi*

(Presented by Academy Member Mindia Salukvadze)

ABSTRACT. The purpose of the paper is to represent a language model that ensures effective functioning of the model in the interactive mode. Interactive approach acquires particular importance in the context of the most complicated processors, as it promises to show how algorithms are simplified, while retaining high level of efficiency. This, for its part, conditions the possibility of using the models of the earlier stages of their development, namely, in the process of their improvement. For example, the possibility of communication with the user while analyzing the sentence makes dictionary "unloaded" from additional information that would be necessary for finding the way out of some "hopeless" situation, and at the same time it requires complication of the algorithm in order to use the information appropriately.

In the case of synthesis in the opposite direction in order to avoid such complications interactive approach also acquires fundamental meaning, namely, it can "go-between" the primary germ of the content and its lingual expression and in this way provide the input of utterance synthesis.

The main purpose of the work is to form the means of representation of the interactive language processor. To solve the task we propose the net approach that was already used earlier to write down certain morphologic algorithms. Nets represent simplified versions of graphs. The most important features of the nets represent three labels associated with each node and with both sides of each arc.

In the case of morphologic processor one of them (node) defines some general grammatical category, two others (arc) points to some particular meaning of the category (left label) and the means of its forming (right label). The transformation of the existing net system is given in the article that is limited to reinterpretation of labels, namely, here they reflect the trajectory that is followed by appropriate information in the question answer mode. The source of information is left arc label, the content of the question is defined by the right label, and the final address is defined by the node label. *© 2013 Bull. Georg. Natl. Acad. Sci.*

***Key words**: sentence synthesis, interactivity, morphologic nets.*

1. Interactive Regime of the Language Model Functioning

The origin of Language model theory is given in [1]. The synthetic direction of this model functioning is proposed as a generator of (quasi-) synonymous sentences with input represented by some, supposedly most neutral sentence, the semantics of which should be as near as possible to the common semantic birth of the whole set of all sentences generated in this way.

The tenor of this work is an attempt to make more exact and definite at least the first step of this process that is the choice of its initial input.

It seems that the interactive approach proposed for solution of this problem realizes both crucial aspects of the process: some "freedom" of choice for the user and quite "strict" control of her/his decisions.

The component of this supposed dialogue system produces the "core" structure of Georgian sentence input: the system asks the user questions concerning the meaning (semantics) of her/his intension ("thought") and on the basis of user's answer builds the grammatically correct core structure, based on the concept of Georgian verb super-paradigm [2].

The system is represented by means of duly adopted Morphologic Nets [3].

2. The Source and Target Objects of Information Exchange

In this paragraph we will consider the sources and addresses of information, which should be received, stored and used in the course of supposed interactive process. The most important, external source of this data, which defines the general style of system functioning, is the user (U), which, firstly, supplies the initial input, which should serve as a basis for definition of the "nucleus" of future sentence, and this initiates the processes as a whole; after that the U continues to "feed" the system with all information, which it cannot procure independently; so, U must "name" all other members of the sentence and define all grammatical features, which cannot be defined proceeding from the already received data (e.g. the tense of verb or the number of nouns).

Each act of this "information flow" should be initiated by some "question" proceeding from the system and the character of the "questions" should depend on some supposed peculiarities of the U(ser) and, particularly, on the supposed level of her/his language knowledge and understanding of the system structure and functioning. These essential and valuable (even from some fundamental point of view) details will be touched on later.

Another precious source of information for this (and all other) language model is a dictionary. One of the main principles, to which the system functioning must be subjected, is a "minimalising" of addresses to U(ser) and a single basis for following this principle is "maximal" use of the internal source of information, that is – of dictionary (D). The dictionary, in its turn, must be sufficiently "rich" in corresponding information to satisfy the requirements of this principle.

The most essential characteristic of the dictionary supposed as a basis for the system under consideration is inclusion in it of a morphologic generator (PG-paradigm generator), which generates full paradigm of the addressed lexical unit.

Another meaningful feature of this dictionary is addition of special units, which correspond to verbal super-paradigm (SP), which imply the unity of verbal paradigms derived from one and the same lexeme. Detailed consideration of this concept in the context of the Georgian language is given in [2]; only the content of such SP units –SPU will be underlined here:

$$\text{SPU: LXF}|\,\text{T}\,|\,\text{P}_1|\,\text{P}_2|\ldots|\,\text{P}_n \qquad (2.1)$$

The symbolic of (2.1) means: **LXF** is the word form or its part (stem, root), which represents this SP as its "head", and serves for identification with this unit of its arbitrary member, that is verb form belonging to one of its paradigms; **T** represents the type of SP, which characterizes the peculiarities of the given SP and specifically its "deviations" from the "regular" (r) SP structure (in what follows we shall imply this "regularity" only); $\mathbf{P}_i$ are pointers to the dictionary units, which represent the verb paradigms belonging to the given SP.

Naturally the assignments of these parameters are constant for each SP unit. Unlike this, the characteristics of output sentence members vary in the course of the process. For example, the current information block (B) of the sentence "nucleus", that the verb which expresses the predicate (PR) on semantic level of structure will be represented by the following set of parameters

PR: SP | DU | WF| VC | M | S |A | T | RW | PS_j | N_j (2.2)

where parameters have the following meaning: **SP** – is a pointer, which fixes the position in the dictionary, where the SP unit is placed, to which the verb belongs; **DU** – analogously points to the dictionary unit, which corresponds to the paradigm which includes the given verb form; **WR** – is assigned by verb forms, which change during the output verb form, which represents the PR on the surface level; **VC**, **M**, **A**, **T** – parameters represent the usual grammatical categories of verb: Voice, Mood, Aspect and Tense respectively.

These categories may have the following set values: **VC** – voice, c(ausative), a(ctive), p(assive) (by that a and p values may be additionally marked by "+" symbols, which mean the version explicitly addressed by affixes: a^+, p^+ (e.g. *uk'etebs* – makes for him/her, *uk'etdeba* – is being made for him/her)); **M** – n(arrative), c(onjunctive), i(mperative); **A** – p(erfect), i(mperfect); **T** – pr(esent), p(ast), f(uture);

The **PR** block includes also some features peculiar for the Georgian language: **S** – series with values: I, II, III; **RW** – row (*mc'k'rivi*), enumerated by numbers 1 to 12.

The value of the category depends on the values of preceding ones (**M**, **A**, **T**, **S**) but for that decisively restricts the area of search for the finally required verb form; the only additional data, which are necessary to fix finally its position in the frames of the whole paradigm are:

$\mathbf{N_i}$, $\mathbf{PS_i}$ – number and person categories, characterizing the actants (one or two), to which the verb addresses explicitly by its affixes (**N** – s(ingular), p(lural), **PS** – 1, 2, 3).

The noun blocks, which correspond to the verb actants (and the end – to the semantic roles (**SR**)) have a far simpler structure:

SR_i: DU | WF| N | C |PS (2.3)

where **DU** points to the corresponding dictionary unit; N, C, PS – define the number, case and person of the noun; the values of N and PS are the same as in (2.2) block; as to case (C) its value spectrum is somewhat larger: n(ominative), g(enitive), d(ative), e(rgative), i(nstrumental), c(ircumstantial). The e, i, c symbols correspond to the following terms accepted in Georgian grammar: *motxrobiti, mokmedebiti, vitarebiti*.

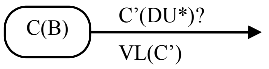
The number and meaning of SR_i and corresponding informational blocks depend on the type of SP. Particularly, regular SP control four SR with their paradigm affixes: causer – CS, agent –AG, object – OB and addressee – AD. Each single paradigm of regular SP chooses one or two of these SR_i to address them explicitly with its affixes. The choice depends on the value of VC category: so, causative addresses CS and AG, active –AG and OB or AD, passive –OB only or OB and AD.

Taking into account that SR_i designation is more "felicitous" in some context (e.g. in the case of cyclic procedures), we will use the double symbolic for the names and pointers of these blocks:

SR_0 | CS, SR_1 |AG, SR_2 | OB, SR_3 |AD (2.4)

As a result, this paragraph defines the following operands for operators, which will be considered in the next section: two sources of information (U(ser) and D(ictionary) and five blocks for accumulation of the current data (PR, SR_0 | CS, SR_1 | AG, SR_2 | OB, SR_3 |AD). Supposedly, the data gathered in these blocks should be sufficient for production of the sentence core structure that is with the verb and its actants.

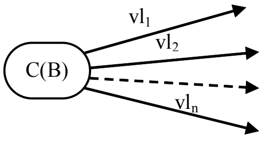
This kind of symbolic can be used also to supply the possibility of assignment to some components of current information blocks (C(B)) the value of some dictionary unit component C'(DU*). Such an analogue of (2.2') operator will look like:



(2.2'')

Remarkable characteristic of operators (2.2), (2.2'), (2.2'') is the fact that they somewhat violate the above-proposed general scheme: particularly, their right arc label (RAL) represents the output, the result of this acts, which should be immediately assigned to the component marked by the node label (NL) (instead of data, which should serve as an input for the object marked by the left arc label (LAL), as it takes place in the rest of the operators considered above (2.1), (2.3), (2.4)).

The operator realizing the choice between possible continuations of the process, which depends on the concrete value (vl_2) of some block component C(B) demonstrates a more radical contrast to the general scheme.



(2.5)

This discrepancy between general scheme (2.5) operator interpretations of the labels "semantic roles" is based on the main orientation of the former on procuring and assignment of information to some block components, on the one hand, and on the obvious purpose of the latter to organize the structure of the process itself, on the other.

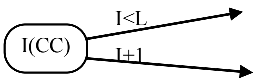
In spite of different character of the process and their final aims, both versions of net representation have much in common. In the first instance, they have identical formal structure; nodes coming out of the arcs and trios of labels (NL, LAL, RAL) marking the node and both sides of the arc. Besides this, the semantics of both means of representation has many points of coincidence, at last on the sufficiently high (perhaps, even metaphorical) level of abstraction.

From this "high" point of view, the processes, which take place in both structures, may be considered as instructions between four elements: one pair of them is participants of a "dialogue and another pair of them represents "messages" which "move" between the pair of dialogue participants: one can be interpreted as a "question" and the other as an "answer" initiated by this "question".

It would be superfluous to repeat the dependence of the above defined operators (2.1)-(2.4) on this scheme. Thus we shall try to show what relation to it may be ascribed to the morphologic net (MN) only. Alike the above-defined operators all three labels of MN are included in some kind of "dialogue", on acts of which is based the whole process of morphologic activity in the frames of language model; the fourth (constantly implied) participant of these acts is the word form, which is synthesized or analyzed. The meaning of the labels is also unchangeable: NL expression some (grammatical) category, as a rule, the particular value of which is represented by LAL and expression of the latter (LAL) in the context of the current word form is given by RAL. Nevertheless, the functions of these participants in the course of "dialogue" acts depend on the direction of processing, that are different for the synthesis or analysis of word form. In terms of the metaphor already used above during the synthesis the word form being produced "asks" the category (NL), what should be the current step of its development, in "answer" NL- category seeks the arc with LAL identical to its current value and proposes to the word form morphologic transformation (e.g. addition of some affix) represented by right label (RAL) of the same arc. In the case of the opposite direction (analysis) this scheme is "turned up": "question" issues from category "answer" from word form, its final content is the required value of NL-category (LAL) chosen as according to the possibility of unification RAL with the word form (in the simplest case - the word form should include an affix expressed by RAL in the corresponding to this affix position). In both cases the "lucky" arc which satisfies the conditions, becomes the continuation of the net process.

Just the lack of such “conditions” in functioning of (3.1)-(3.4) operators make difference between them and morphologic nets: the former always imply the “unconditional” receiving and assignment of information. The (3.5) operator must fill up this gap and so supply the possibility of choice among different branches of net structure, which is the basic feature of the net representation as such.

Lastly, the “economic” structurization of the net representation requires addition of more operators, which supplies the possibility of cyclic procedures:

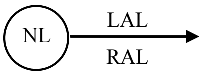


(2.5’)

The (2.5’) operator implies existence of an additional block (CC) with two components (I, L), which represent the variable **I**ndex and its **L**imit respectively; the use of (2.5’) implies two preceding operators of (2.2) type, which assign to I and L the initial and final values of the variable I organizing the cycle (CC); the part of net subordinate to this cycle operator (2.5’) is repeated till the moment, when I gets equal to L.

3. The Net Represented Operational System

General semantic of the net scheme



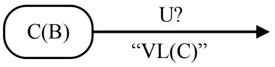
N(ode), L(abel), L(eft) A(rc) L(abel), R(ight) A(rc) L(abel)

The main scheme of semantic relations between these L(abels) is: NL asks the LAL what is its (NL’s) value (RAL); RAL defines this value and assigns it to NL.

So, we have something like, dialogue between NL and LAL, which designs by the question from NL to LAL, and ends by the answer of LAL, which satisfies the NL’s requirement.

Naturally, the “question” of NL, that is LAL, must contain sufficient information to define for LAL, what “answer” is expressed from it.

These semantic relations between L(abels) are most obviously demonstrated by the central link of the interactive process, that is by the act of dialogue between the system and its user which is given below as the first (3.1) example: NL=C(B), LAL=U?, RAL=VL(C)

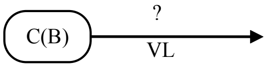


(3.1)

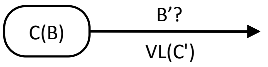
U(ser), B(lock) of the current information Value of C(omponent/category).

The question “VL(C)” is addressed to U: U’s answer should define VL of C(B), which will be assigned to C(B).

The value (VL) of C can be defined and assigned without address to some external source (U). The simplest version of such assignment is immediate demonstration of VL in the operator context:



(3.2’)



(3.2)

As a result VL-value will be immediately assigned to C(B). The value of C may also be copied from some other (C’) component of the same or another information block (B/B’).

One more version of C(B) values (VL) definition is the address to the dictionary (D), which represents the most important sub-set (SS) of this interactive system as a whole, and at the same time, is its most internal component (and as such is completely opposed to the U(ser)).

The proposed scheme of the sentence interactive synthesis is based, particularly, on the Georgian Computer Dictionary, created in the frame of the Project, supported by Rustaveli Foundation and fulfilled at the Institute of Control Systems (2009-2011). The specificity of this quite voluminous (nearly 100 000 units) dictionary is represented by inclusion in it of the Morphologic Processor, that is of the Morphologic Generator, which defines for each dictionary unit the whole list of its paradigm members. As a result of such combination, this dictionary (unlike the usual ones) represents the unity of all word forms of language (and not all the lexemes only); and because of this it can "answer" by the concrete word form, if the question contains sufficient information for the choice of the needed form from the whole list of the corresponding paradigm members.

Thus the address to the dictionary can be interpreted like a two step procedure: first of them implies the definition of the corresponding dictionary unit (DU) position in the dictionary; the other can be initiated by the "order" to generate the paradigm corresponding to the given unit and single out the needed member of this paradigm. The input of the former should be some (arbitrary) word form belonging to this unit paradigm; as for the latter its input should contain the whole information necessary for the choice of the required form out of the whole set of paradigm members. The word form chosen so should be returned by the second step and copied in corresponding block component.

Taking into account that the dictionary unit includes, besides its paradigm, many other necessary data, it seems sensible to fix the position of the dictionary unit already identified or the first step for the further addresses to this information (without excessive operational expenses and "bothering" of the user). So, we will suppose that the first step of addresses to the dictionary returns the pointer (PN) at the identified dictionary unit, which becomes the value of corresponding component of some block (B).

The difference between inputs/outputs of the above mentioned steps of procedure justifies their implementation by means of different operators with the only restrictions: the activation of the first step operator should precede all other addresses to the dictionary unit fixed by it.

The first step, supplying the definition of dictionary unit to which belongs the given word form, can be performed by the following operator:

$$\boxed{\mathrm{DU(B)}} \xrightarrow[\mathrm{WF}]{\mathrm{D?}} \qquad (3.3)$$

$$\boxed{\mathrm{SP(PR)}} \xrightarrow[\mathrm{WF}]{\mathrm{SPL?}} \qquad (3.3')$$

According to many considerations, it seems sensible to single out the set of verbal super-paradigms and represent it as a separate list of SP units (SPL). Taking into account that this list should be addressed by the predicate (PR) block only, we can realize this operation by (3.3').

Now the system can - on the basis of information procured by (3.3) - apply to dictionary units defined so and "require" from them the members of their paradigms which represent just the word forms necessary in the context of the final output sentence. More exactly, this address should be made to the paradigm generating (PG) components of these units and must be accompanied by the grammatical data fixing the position of output word form in the list of generated paradigm members. Naturally, this bunch of grammatical features is different for each class of word forms and, in the first instance for the different parts of speech: operators (3.4) and (3.4') demonstrate this peculiarity as an example by verbs which express the predicate, and nouns realizing all other semantic roles (SR_i) included in the set of blocks mentioned in the previous paragraph:

$$\boxed{\mathrm{WF(PR)}} \xrightarrow[\mathrm{RW+PRS_1+N_1+PRS_2+}]{\mathrm{PG(DU^*)?}} \qquad (3.4)$$

$$\boxed{\mathrm{WF(SR)}} \xrightarrow[\mathrm{N+C}]{\mathrm{PG(DU^*)?}} \qquad (3.4')$$

According to (3.4), (3.4') it is necessary for definition of output verb form to supply information about its row (RW) and about grammatical features characterizing its actants (not more than two), to which the verb explicitly addresses by its affixes. Unlike this, it is sufficient in the case of noun to mention the values of its number (N) and case (C) categories.

One more peculiarity of (3.4), (3.4') is the use of affixes (*) as the upper index of DU. These (usual for some programming systems) symbols underline the fact that the right arc label (RAL) of these operators implies the unit itself, but not the content current value of DU component, which in reality is only the pointer orienting this dictionary unit and defining its position in the list of all dictionary units.

Thus, we have considered the means of net representation system which is used for the creation of an algorithm of the Georgian core structure synthesis.

ენათმეცნიერება

ინტერაქტიული ენობრივი პროცესორის ქსელური წარმოდგენა

გ. ჩიკოიძე*, ა. ჩუტკერაშვილი*, ნ. ჯავაშვილი*

* *საქართველოს ტექნიკური უნივერსიტეტის ა. ელიაშვილის მართვის სისტემების ინსტიტუტი, თბილისი*

(წარმოდგენილია აკადემიკოს მ. სალუქვაძის მიერ)

ნაშრომის მიზანია ენობრივი მოდელის წარმოდგენა, რომელიც უზრუნველყოფს მოდელის ეფექტიან ფუნქციონირებას ინტერაქტიულ რეჟიმში. ინტერაქტიული მიდგომა განსაკუთრებულ მნიშვნელობას იძენს ურთულესი ენობრივი პროცესორების კონტექსტში, რადგან ის გვპირდება ალგორითმების გამარტივებას მათი შედეგიანობის მაღალი დონის შენარჩუნებით. ეს კი, თავის მხრივ, განაპირობებს მოდელების გამოყენების შესაძლებლობას მათი განვითარების შედარებით ადრინდელ ეტაპზე, კერძოდ, მათი სრულყოფის პროცესში. მაგალითად, წინადადების გაანალიზებისას მომხმარებელთან კავშირის შესაძლებლობა ლექსიკონს "განტვირთავს" იმ დამატებითი ინფორმაციისგან, რომელიც აუცილებელი იქნებოდა ზოგი "ჩიხური" სიტუაციიდან გამოსასვლელად და, ამავე დროს, მოითხოვს ალგორითმის სათანადო გართულებას ამ ინფორმაციის სათანადო გამოყენებისათვის.

საპირისპირო, ანუ სინთეზური მიმართულების შემთხვევაში, ანალოგიური "გართულებების" თავიდან აცილების გარდა, ინტერაქტიული მიდგომა ფუნდამენტურ მნიშვნელობასაც იძენს, სახელდობრ, მას შეუძლია "იშუამავლოს" შინაარსის პირვანდელ ჩანასახსა და მის ენობრივ გამოხატულებას შორის და ამგვარად უზრუნველყოს გამონათქვამის სინთეზირების შესავალი.

უშუალოდ ნაშრომის მიზანია ინტერაქტიული ენობრივი პროცესორის წარმოდგენის საშუალებათა ჩამოყალიბება. კერძოდ, ამოცანის გადასაწყვეტად ნაშრომში შემოთავაზებულია ქსელური მიდგომა, რომელიც ადრე იქნა გამოყენებული რიგი მორფოლოგიური ალგორითმების ჩასაწერად. ქსელები წარმოადგენენ გრაფის გამარტივებულ ვარიანტს; ქსელების ყველაზე მნიშვნელოვან მახასიათებელს წარმოადგენს ჭდეების სამეული ასოცირებული ყოველ კვანძთან და ყოველი რკალის ორივე მხარესთან.

მორფოლოგიური პროცესორის შემთხვევაში ერთ-ერთი მათგანი (კვანძი) განსაზღვრავდა რომელიმე ზოგად გრამატიკულ კატეგორიას, დანარჩენი წყვილი (რკალი) მიუთითებდა ამ კატეგორიის რომელიმე კონკრეტულ მნიშვნელობაზე (მარცხენა ჭდე) და მისი გაფორმების საშუალებებზე (მარჯვენა ჭდე). მოცემულ ნაშრომში შემოთავაზებულია აგრეთვე არსებული ქსელური სისტემის ტრანსფორმაცია, რომელიც შემოიფარგლება ჭდეების რეინტერპრეტაციით: სახელდობრ, ისინი აქ ასახავენ ტრაექტორიას, რომელსაც მიჰყვება გარკვეული ინფორმაცია "კითხვა-პასუხის" რეჟიმში: ინფორმაციის წყაროს წარმოადგენს რკალის მარცხენა ჭდე, კითხვის შინაარსს განსაზღვრავს მარჯვენა ჭდე, მის საბოლოო მისამართს კი — კვანძის ჭდე.

REFERENCES

1. *I. Mel'chuk* (1974), Opyt teorii lingvisticheskikh modelei "Smysl↔Tekst", Moscow (in Russian).
2. *G. Chikoidze* (2010), Sistematizatsiia znachenii nekotorykh klassov iazykovykh edinits. Monograph. Tbilisi (in Russian).
3. *G. Chikoidze* (2004), Setevoe predstavlenie morfologicheskikh protsessorov. Monograph. Tbilisi (in Russian).

Received December, 2012