STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **XLVI**, Number 2, 2001

ON ROMANIAN ARTICLE SEMANTICS

DANA AVRAM

Abstract. In this paper a way to represent Romanian article semantic is proposed. When trying to represent sentence semantics by using first order predicate calculus, the articles usually became the mathematical operators $\exists$ and $\forall$. We define a more powerful operator, called DET, that encloses the significance of $\exists$ and $\forall$. Rules that may be used with DET in a FOPC are considered. This proposed representation is also appropriate for other determiners class.

1. Introduction

The kinds of grammars that we are familiar with do not model the reasoning process of the brain. They are descriptions of the natural language structure to a certain degree of precision. So the course of development of the brain cannot really be explained by the formal properties of the descriptive apparatus. Similar questions can be raised for every piece of knowledge, specific and general, that is embodied in the mental grammar. The set of things that we characterize as prior knowledge are part of the language faculty, or universal grammar [2]. Everything else is learned. The question, now, is how?

Deduction systems are one of the most used expert systems. They have a set of statements (dates) and a set of deduction rules for constructing new statements. They are built on mathematical logic operations whose fundamental rules are very well stated. This very rigid but rigorous system will be exploited in the next paragraphs.

2. Articles in Romanian Language

Syntactically speaking, the article appears near a noun [5]. They are definite, indefinite, or bare. The indefinite form usually indicates that the speaker has no information, or a reduced degree of information, on the object it stands by. The definite article indicates a higher degree of information. The bare article is an indefinite article, but its presence is not required by the syntactic rules. The singular and plural form, the definitiveness or indefinitiveness article forms have a

2000 *Mathematics Subject Classification.* 68T50.

1998 *CR Categories and Descriptors.* I.2.7[**Computing Methodologies**]: Artificial Intelligence – *Natural Language Processing*.

supplementary numeric role. We will concentrate on the second problem and will construct a model that captures this meaning in a simple and explicit way [4].

From the viewpoint of a syntactic form, the article may exist by himself as a separate word, standing above the noun, or it may be part of the noun it determines. The forms of the Romanian article are presented in Table 1.

TABLE 1. The Romanian article syntactic forms

	Article	**Syntactic form**	**Number/Use**	**Example**
1.	*-l,-le,-a -lui, -e, -i*	part of the noun	singular/definite reference	*cainele [the dog]*
2.	*un, o, unei, unui*	separate word	singular/indefinite reference	*un caine [a dog]*
3.	*(toti) -i,(toate)-le-lor*	separate word (optional) and part of the noun	plural/definite reference	*(toti) cainii [(all) dogs]*
4.	*(niste) -le, -lor*	separate word (optional) and part of the noun	plural/indefinite reference	*niste caini [some dogs]*
5.	BARE		bare singular NP (mass term)	*caine [dog]*
6.	BARE		bare plural NP (generics)	*caini [dogs]*

3. Representation using FOPC

Let $L = (\Sigma, F, A, R)$ be the first order predicate logic (FOPC):

$$\Sigma = V \cup C \cup (\cup F_j) \cup (\cup P_j) \cup \{\forall, \exists, \neg, \wedge, \vee, \rightarrow, (,)\},$$

where V represents a set of symbols called variables, C represents a set of symbols called constants, F_j represents a set of function symbols with j parameters, P_j represents a set of function symbols with j parameters, P_j represents a set of predicate symbols with j parameters, the set $\{\forall, \exists, \neg, \wedge, \vee, \rightarrow, (,)\}$ represents the logic operators, and $\exists$ and $\forall$ represent FOPC quantifiers.

One important class of semantic constructors is quantifiers class [3]. In the first order predicate calculus [8], the two quantifiers, $\forall$ and $\exists$, encode the articles' meaning.

Let us see how we can represent some sentences using the FOPC defined earlier.

The singular article, definite or indefinite (Table 1, lines 1 and 2), says that $\exists$ the material object that corresponds to the given noun.

The definite plural article (Table 1, line 3) has the meaning of $\forall$.

The indefinite plural article (Table 1, line 4) indicates the existence ($\forall$) of one or more objects of the type that corresponds to the given noun. Few more quantifiers are necessary.

The bare (missing) article (Table 1, lines 5 and 6) is usually interpreted as indefinite article.

TABLE 2. Examples of sentences representation using FOPC

	Romanian	**English translation**	**Representation of Romanian sentence**	**The same sentence in English**
1.	*Un caine latra*	*[A dog barks]*	($\exists x$: (CAINE(x)) $\rightarrow$ LATRA(x))	($\exists x$: (DOG(x)) $\rightarrow$ BARK(x))
2.	*Toti cainii latra*	*[All dogs bark]*	($\forall x$: (CAINE(x)) $\rightarrow$ LATRA(x))	($\forall x$: (DOG(x)) $\rightarrow$ BARK(x))
3.	*Niste caini latra*	*[Some dogs bark]*	($\exists x$: (CAINE(x)) $\rightarrow$ LATRA(x))	($\exists x$: (DOG(x)) $\rightarrow$ BARK(x))

Looking at Table 2 we may remark that the sentence *Un caine latra* / [*A dog barks*] is represented in the same way as the sentence *Niste caini latra* / [*Some dogs bark*]. It is easy to notice that this representation looses a part of the natural language semantics. The problem is that any natural language contains a much larger range of quantifiers than the two from FOPC. As an example for the higher complexity of natural language quantifiers, let us consider the following FOPC formula

$$\forall x : P(x)$$

This formula is true if and only if $P(x)$ is true for every possible object in the domain.

Such statements are rare in natural language. We will rather say [*most dogs bark*] (and in this case, this is not an article domain, but a Romanian adverb) or [*some people laugh*], which requires constructs that are often called generalized quantifiers. These quantifiers are used in statements of the general form [1, 6]:

(quantifier variable: restriction proposition $\rightarrow$ body-proposition)

For example:

([NISTE](x):(CAINE(x)) $\rightarrow$ LATRA(x))
(SOME(x):DOG(x) $\rightarrow$ BARK(x))

This roughly captures the meaning of the sentence: *If there are some things that are also dogs, then they are barking things.*

Or:

([CEI MAI MULTI](x):(CAINE(x)) → LATRA(x))
(MOST(x):DOG(x) → BARK(x))

This means that: *Most dogs are barking things.*

4. Quantifiers with extended functionality

A construct to handle plural forms, as in the phrase *two dogs bark* must to be introduced. This indicates not *a dog*, but *two*. It can easily be seen that the article has also a numeric meaning. Let us consider the general form:

DET[*variable*, *name*, *number*]

where *variable* is the variable inherited from FOPC, *name* is the name of the determinant, and *number* is the number of objects indicated by the noun, or the percent value (of all the possible objects in discussion) only indicators of number (as *all, some*) are specified. *Toti* [*all*] refer to 100%, some will be convenient for 25% (less than 50%). Tabel 3 shows the Romanian articles representations using the general form DET described above.

Table 3. Articles representations using DET

	Article	**Representation**
1.	*-l,-le,-a-lui, -e, -i*	DET[x, **article**, 1]
2.	*un, o, unei, unui*	DET[x, **article**, 1]
3.	*(toti) -i,(toate)-le-lor*	DET[x, **article**, 100%]
4.	*Niste -le, -lor*	DET[x, **article**, 25%]
5.	BARE	The same representation as the indefinite form and the same number

For example, the previous expression

([NISTE](x):(CAINE(x)) → LATRA(x))

becomes

(DET [x, **niste**, 25%]:(CAINE(x)) → LATRA(x))

and the expression

([UN](x):(CAINE(x)) → LATRA(x))

becomes

(DET[x, **un**, 1]:(CAINE(x)) → LATRA(x))

For the adverbial expression *cei mai multi* [*most*] 75% will be convenient for (more than 50%).

For example, the expression:

([CEI MAI MULTI](x):(CAINE(x)) → LATRA(x))

becomes:

(DET [x, **cei mai multi**, 75%]:(CAINE(x)) → LATRA(x))

Numeral determiners make no assumption about the whole class of the object. Their number will appear on the third position on DET argument, as in the next example.

The expression:

([DOI](x):(CAINE(x)) → LATRA(x))

becomes:

(DET [x, **doi**, 2]:(CAINE(x)) → LATRA(x))

5. Rules for DET

To allow quantifiers, variables are introduced as in first order logic but with an important difference. in first order logic a variable only retains its significance within the scope of quantifier. Thus two instances of the same variable x occurring in two different formulas – say in the formulas $\exists x P(x)$ and $\exists x Q(x)$ are treated as completely different variables with no relation to each other. Natural languages display a different behavior. For instance consider that two persons say the following two true sentences: *A dog barks* and *Three dogs bark*. The first sentence introduces a new object to the discussion namely *a dog*. You might think to treat the meaning of this sentence along the lines of the existential quantifier in logic. But the problem is that the dog number introduced existentially in the first sentence is completed by to the number *three* in the second sentence. Variables appear to continue their existence after being introduced and the associate determiners usually change by unification [1]. In FOPC they are combined using the logical operators $\{\neg, \wedge, \vee, \rightarrow, (,)\}$. Here $\rightarrow$ can be obtained from $\vee$ and $\neg$, and the parens (,) specify the order. So we have to consider rules of combining DET with $\rightarrow, \wedge, \vee$.

We suppose that all the variables refer to the same variable univers (unique and known) which will be called the contextual universe.

Rules for $\neg$ are different when the number is percent and when it has a concrete value. As it will be seen, there are cases when taking a decision is improper.

Rule no 1: (Percent case)

$\neg$((DET [x, **det1**, p1]: OBJECT(x)) = (DET [x, **det1**, 100% -p1]: $\neg$ OBJECT(x))

Rule no 2: (Numeric case)

Suppose that we know the total number objects in the contextual universe.

if **tot** = total_nr_of_objects_that_determiner_determines is known then

$\neg$ (DET [x, **det1**, p1] : OBJECT(x))= (DET [x, **det1**, **tot** -p1] : $\neg$ OBJECT(x))

else

$$\neg\ (\text{DET}\ [x,\ \textbf{det1},\ p1]\ :\ \text{OBJECT}(x)) = \text{undefined}$$

The rule for $\wedge$ depends on maximal values of numeric argument. Suppose that $p1 > p2$.

Rule no 3:

$$((\text{DET}\ [x,\ \textbf{det1},\ p1]\ :\ \text{OBJECT}(x)) \wedge (\text{DET}\ [x,\ \textbf{det2},\ p2]\ :\ \text{OBJECT}(x))) = (\text{DET}\ [x,\ \textbf{det1},\ p1]\ :\ \text{OBJECT}(x))\ (\text{if}\ p1 > p2)$$

Rule for $\vee$ depends on numeric argument as in $\wedge$ case.

Rule no 4:

$$((\text{DET}\ [x,\ \textbf{det1},\ p1]\ :\ \text{OBJECT}(x)) \vee (\text{DET}\ [x,\ \textbf{det2},\ p2]\ :\ \text{OBJECT}(x))) = (\text{DET}\ [x,\ \textbf{det2},\ p2]\ :\ \text{OBJECT}(x))\ (\text{if}\ p1 > p2)$$

There are mathematical rules that link with $\exists$ and $\forall$. One of the simplest mathematical rules [8] says that $\forall$ implies $\exists$.

if

$(\forall x\colon \text{OBJECT}(x))$

then

$(\exists x\colon \text{OBJECT}(x))$

That means that: if *any x is object* is true, then *an x is object* is true, too. In the DET case that is:

Rule no 5:

if

(DET $[x,$ **det1**, $p1]$: OBJECT $(x))$ and $p2 < p1$ (percent or numeric)

then

(DET $[x,$ **det1**, $p2]$: OBJECT $(x))$.

This means that if there are $p1$ objects x and $p2$ is such that $p2 < p1$, then there are also $p2$ objects x (in the given contextual universe).

We saw that problems that appear in the percent case are solved if we know the total number of the objects in the contextual universe. This is so because in this case we can transform the percent value into a (real) numeric one, as in the rule that follows.

Rule no 6:

Suppose that **tot** is the total number of the objects OBJECT in the contextual universe and $p1$ is a percent value. Then the math says that: $p2 = p1/100 \times \textbf{tot}$, and $p2$ is a numeric value, i. e.:

(DET $[x,$ **det1**, $p1]$: OBJECT $(x))$ = (DET $[x,$ **det1**, $p1/100 \times$ **tot**] : OBJECT $(x))$.

Taxonomies are valuable resources in Natural Language Processing and Artificial Intelligence. They consist of hypernym (generalization) and hyponym (specialization) relations between concepts [7]. The most known example of such an organization is WordNet – a lexical database organized as a general terminological

system that contains semantic classes organized hierarchical is a classical example. There are also other knowledge bases as a partially structured knowledge, as domain specific terminological systems. Those structures may be easily used, if available, to improve deduction rules into determiners domain.

Let us consider the case of hierarchical system generated by ISA arcs. Suppose that OBJECT1 ISA OBJECT2, like a DALMATIAN is a DOG.

Rule no 7:

if

(DET [x, **det1**, $p1$] : DALMATIAN (x)) and $p1$ is numeric

then

(DET [x, **det1**, $p1$] : DOG (x))

This deduction rule works as follows: if there are five Dalmatians that bite, then there are also (at least) five dogs that bite.

This rule does not work in the percent case. We cannot say that if there are 50% Dalmatians that bite, then there are also 50% dogs that bite, nor that if there are 50% dogs that bite, then there are also 50% Dalmatians that bite.

6. Further research

Examples regarding the article case have been discussed. The issues approached in this paper may be developed even for numerals and other adverbs (with determiner role). As we have seen, the article semantics is much the same as of the other parts of speech with determiner role.

Only the case of a unique universe has been considered. But in the real world, each speaker states truths about his own known, time changing universe which may be different from the others. There is always a possibility that the statement be not (exactly) true if reported to the general universe. The classification scheme, structured according to the state of current human knowledge is, also, not perfect. Sometimes, the hierarchy is not so well done, there are exceptions that must be handled. One solution would be to introduce a special parameter to DET argument list in order to handle those special cases.

Noun phrases serve many different language functions, and it is important to distinguish these functions when considering scoping issues. There are at least three major classes to consider. Those involving definite reference indicate that the listener should in principle be able to identify the object or set. Definite reference occurs, for example, with determiners as *the* as in *the dog* (an individual) or *the fat men* (a specific set). In any natural settings there will obviously be many dogs in the world, so the use of the context to identify the correct one is crucial for understanding the sentence. Identification is a problem of anaphora resolution, and has been widely discussed in the literature.

References

[1] **Allen, J.**, Natural Language Understanding, *The Benjamin Cummings Publishing Company, New York*, 1995.

[2] **Culicover, P.W.**, Language acquisition and the architecture of the language faculty, *Proceedings of the Berkeley Formal Grammar Conference Workshop, The University of California, Berkeley, CSLI Publications, http://www-csli.stanford.edu/publications/*, 2000.

[3] **Gal, A., Lapalme, G., Saint-Dizier, P., Somers, H.**, Prolog for Natural Language Analysis, *John Wiley, London*, 1991.

[4] **Graur, Al.** et. al., Romanian Language Grammar, *Romanian Academy Publishing House*, 1966.

[5] **Jurafsky, D., Martin, J.M.**, Speech and Language Processing, *Prentice Hall, Inc., University of Colorado, New Jersy, USA*, 2000.

[6] **Onet, A.**, A module-based application for te semantic representation of natural language sentences, **Proceedings of Eurolan 2001, Iasi, Romania**, 2001.

[7] **Stevenson, M.**, Enriching Noun Taxonomies with Thesaural Information, *Proceedings of NAACL 2001, Carnegie Mellon University Pittsburgh, PA, USA*, 2001.

[8] **Tatar, D.**, Artificial Inteligence: Automate Demonstration, Natural Language Processing, *Editura Albastră*, 2001.

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania
E-mail address: davram@cs.ubbcluj.ro