"FORMULAE" IN COHERENT TEXT : LINGUISTIC RELEVANCE OF SYMBOLIC INSERTIONS

FELIX DREIZIN

Responsa Project
Bar-Ilan University
Ramat-Gan, Israel

Some difficulties in automatic analysis and translation bound to symbolic insertions in mathematical texts are discussed. Rules dealing with these difficulties are proposed. These rules are based on the use of the whole text of the article incorporating a formula.

For satisfactory automatic analysis of texts, it is necessary to provide in the dictionary exhaustive semantical information ascribed to its entries. But this information can appear to be insufficient in cases where the meaning of linguistic elements is ascribed to their occurrences by the very text in which they are encountered cf. for example, pronouns.

The other example is provided by symbolic insertions in mathematical texts, which we shall call "formulae". So not only 'a = b', 'x > y' etc., but also

'x' ' $\sqrt{2}$ ' and so on are "formulae".

Mathematical formula resembles pronouns in one respect: it is semantically "void" being out of context.

For example, 'G' may be "set", "subset", "group", "operator", "function", "string", "element", "rule of grammar", etc.

The meaning is ascribed to a formula by the context. There are a few types of formulae with fixed meanings. For example, 'dx/dy' is 'derivative'. But this situation is not typical.

One of the basic usages of formulae consists of naming by formula A some individual object a belonging to some class b of objects such that there exists some noun block Cl(A) that names b.

For example, in the expression 'set R' the formula 'R' names some individual set belonging to the class of "sets". Noun block 'set' (consisting in this case of a single noun) names this class.

So here

$$Cl(R) = \text{'set'}.$$

Consider some difficulties arising in translation because of the absence in a source sentence of the Cl(f) for a formula f.

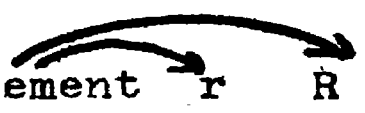
Let us try to translate from English to Russian the sentence

(1)

(Previous surface syntactical analysis is assumed, its results being represented in dependency-tree form).

Syntactically, this sentence is very simple, but even an experienced "human" interpreter would not be able to properly understand this expression and translate it.

In Russian, the element corresponding to the English preposition 'of' is, generally, the grammatical meaning "genitive". We can ascribe this meaning to the formula 'R' (governed by the preposition 'of'):



(Syntactical links are also shown).

At this point, the process of translation is suspended because of the fact that in Russian two non-coordinate formulae cannot depend on the same noun.

Similar examples are provided by other languages:

French: 'L'image X de Y'.

German: 'In jeder Umgebung V von X'.

Armenian: 'X-i x arzhekhə '. etc.

A human interpreter does not usually hesitate to properly translate such expressions only because he understands their meaning from a general background or vast context. We can point out some characteristic constructions in mathematical texts that are sufficient as contexts in such cases. Consider, for example, such a context (i.e. an expression from the same text):

'Let R be a ring with a unity 1'. (2)

for expression (1), and let us formulate very simple rules:

'Let f be a N' → Cl(f) = 'N' ; (R_1)

'Cl(f) f' ← 'f' ; (R_2)

N f' → Cl(f) = N ; (R_3)

Here f is some "formula", N - a noun block, ⌒ means syntactical link, → reads: 'if ... then', and ← means substitutability.

With the aid of the rules R_1 and R_2 we can obtain from (1) and (2) the following expression:

We first find an element r of ring R

which is easily translatable to Russian:

(The relevant syntactical links are shown; the two "formulae" depend on different nouns).

Of course, 'ring R' is not substitutable for R in the expression 'ring R'

The expression

'Components x_i are nonnegative'

with the aid of the rule R_3 provides us $Cl(x_i)$ and helps to translate the expression:

'A unique value a of x_i '

Cf. also the contexts:

'L'espace topologique Y' and

'Ein topogener Raum X'

for the French and German examples above.

Let us now try to translate to Russian the following expressions:

'H is cyclic' (3); 'A smallest k' (4).

A predicative adjective in Russian must be put in grammatical agreement with the subject of the sentence; an attributive adjective - with the governing noun. That is, the Russian adjectives for cyclic' in (3) and for 'smallest' in (4) must agree in gender

with 'H' and 'k' correspondingly. It is clear that the information about the gender of a "formula" can be provided by Cl(f). Having defined Cl(H) = 'matrix' for which the translation

'MATRITSA'

is feminine, we receive for (3) the translation

'H javljajetsja tsiklicheskoj'.

There exist numerous other expressions for which the finding of Cl(f) is very desirable, for example:

We define j and k by j = m + n; k = m - n'. (5)

The "direct" translation of (5) to Russian:

'Opredelim j i k putjom j = m + n; k = m - n'

is not smooth enough; the translation:

'Opredelim j i k s pomoshju sootnoshenij

j = m + n; k = m - n'.

("We define j and k by correspondences j = m + n; k = m - n") is much better.

Cl(f) can be sometimes defined from the very formula f. For example, 'a = b' is "equality", 'a > b' is "inequality", and so on. Sometimes the "meaning" of a formula f can be derived from words syntactically linked to this formula or from a more complex formula F incorporating f.

For example, from the expression

'T : A $\longrightarrow$ B'

we can derive that 'T' is a "transformation" and that 'A' and 'B' are "sets". From the expression

'a $\in$ B'

we can derive that 'B' is a "set" and that 'a' is an "element".

In

'Subset of A'

'A' is a "set". In

'Differentiation (or: integration) with respect to x', 'x' is a "variable", and so on.

Cl(f) for a formula f can be sometimes a more or less bulky expression consisting of a noun with words depending on the noun directly or indirectly.

Example:

'Tous les ensembles L_i d indices inferieurs a un nombre donne K' (Cl(L_i) is underlined). (6)

In this case we can reduce Cl(L_i) to only one word 'ensembles'. But in rare cases such reduction will produce absolutely inadequate translations.

Example:

In the expression

'Pour les fonctions x(t) de L' (7)

with a context

'La partie commune L de tous les ensembles L_i'

(Cl(L) is underlined),

we cannot reduce Cl(L) to only one word 'partie', which is its syntactical governor.

It is very difficult to formulate a general rule to discriminate between cases of types (6) and (7).

The expression (7) can be translated using a synonym for Cl(L), for example, 'ensemble', having in mind that the intersection of several sets is also a set. The computability of such synonyms can, of course, be questioned.

* *
*

Now we shall consider some problems arising in translation of Russian mathematical texts into European and other languages.

The construction

in Russian has two syntactical meanings,

(a) appositive:

'podmnozhestvo B'
= "subset B";

(b) genitive:

'podmnozhestvo B'
= "subset of B".

The cause of this difficulty is the omission of Cl(f) in the surface syntactical structure of some Russian sentences:

'podmnozhestvo mnozhestva B' ⟶ 'podmnozhestvo B'

"Subset (of) set B" "Subset (of) B".

In such cases the genitive link is rare (5% of all occurrences of constructions of type N ⌒ f, i.e. several-dozen occurrences in a mathematical article). The task of automatic choice here is very difficult. It was solved only partially. We can choose from the text of an article about 70% of all occurrences of the appositive links and also some occurrences of genitive links. The rest of occurrences remain ambiguous.

The proposed procedures were checked in exhausting manual experiments, but their adaptation for computer is quite feasible.

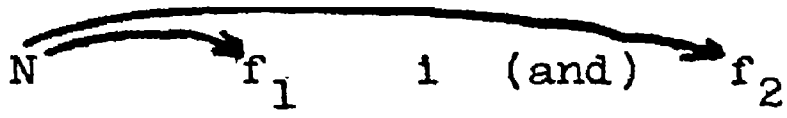
Choice of appositive links

Let us consider the following empirically stated axioms to be valid:

(A_1) In the same Russian text every two different occurrences of the same expression of type N ⌒ f are or both appositive or both genitive.

So, if we have succeeded in clarifying the meaning of a link in one occurrence of a construction, we can ascribe this meaning to every occurrence of the same construction.

(A_2) In a construction of the type



where f_1 and f_2 are two syntactically co-ordinate formulae, the two links are both appositive or both genitive.

For example, having a construction

'mnozhestva A i B'
("sets A and B" or "sets of A and B")

and knowing that in

'mnozhestva A'

the link is appositive, we can consider the link in

'mnozhestva B'

to also be appositive (i.e. "sets A and B").

(A_3) Let us call constructions of the type

'f jest' N'
("f is N")

and

'oboznachim N cherez f'
("Let us designate N by f")

introductory constructions. Every introductory construction ascribes the meaning to the formula which it introduces.

> In every construction $N \overset{\frown}{\quad} f$, for which an introductory construction exists in the same text, the link is appositive

(A_4) Sometimes the meaning is ascribed to a formula without any introductory construction.

> The link in an occurrence r of a construction of type $N \overset{\frown}{\quad} f$ is appositive, if the expression f has not occurred in the text before r.

In this case the formula f must also not occur before r as any coherent part (subformula) of some other formula F, because the meaning can be ascribed to a formula f by its place in F (see above). But to use the distinction between a coherent and a non-coherent part of a formula (Cf. 'a + b' in '(a + b)/d' and in 'ca + bd'), we need a calculus of all mathematical symbolic notations, of which only small portions exist (Cf. arithmetic expressions of programming languages). Because of this A_4 was formulated in the above form,.

(A_5) Sometimes there occur in mathematical texts expressions where verbal and symbolic parts are interwoven

so that in syntactic analysis a symbolic insertion appears not as a single unit but as a complex construction having its own structure. Some parts of a formula can have links of their own with the external verbal parts of the sentence.

Examples:

1. 'funktsija $L \in H(R)$'.
("Function $L \in H(R)$").

Here ' $\in$ ' is the predicate of the sentence, 'Function' is its subject and 'H(R)' - an indirect object. The sentence can be read as 'Function L belongs to H(R)'.

2. 'Dlja ljubogo $l \in B$ imejem' ...
("For every $l \in B$"...)

Here '$\in$' is an attribute of l and can be read as 'belonging to'.

3. 'funktsija $L \in H(R)$ opredeljajetsja'...
"Function $L \in H(R)$ is defined by"...)

Here '$\in H(R)$' is an attribute of 'Function', and 'L' is an apposition modifying the same word. But the whole string $L \in H(R)$ can also be considered an apposition modifying the word 'function'. So, we can formulate a rule: If the link in some construction of the type N $\overset{\frown}{\quad}$ f is appositive, then the link of the same N with the formula f R f', where R is one of the symbols $=, \neq, <, \leq, >, \geq, \supset, \supseteq, \subset, \subseteq, \in, \equiv$ or $\not\equiv$ and f' is a (coherent) part of the

formula f R f' is also appositive.

The inverse also holds true.

Using the axioms A_1 to A_5 cyclicly, we receive the 70% mentioned above.

Example:

Let us assume that the following Russian expressions belong to the same mathematical text (and the preliminary syntactical analysis has already been done):

(1) 'Oboznachim etu tsepochku cherez A'
("Let us designate this string by A");

(2) 'tsepochki A i B = D'
("strings. A and B = D"? Strings of A and B = D"?);

(3) tsepochki B i F'
("strings B and F"? "Strings of B and F"?);

(4) tsepochka F'
("string F"? "String of F"?)

Using axioms A_3, A_1, A_2, A_5, A_1, A_2 and A_1 we can ascribe the meaning "appositive" to the link in (4).

Assuming that in a text expressions (2), (3) and (4) are present, and that the occurrence of 'F' in (4) is the first occurrence of this formula, we can ascribe to the

link

'tsepochki A'

in (2) the meaning "appositive" with the aid of axioms A_4, A_1, A_2, A_5 and A_2.

So, we receive for expressions (2) to (4) translations: "strings A and B = D"; "strings B and F"; "string F"

It is worthwhile to mention that the same formula may occur in a text being linked appositively to several different nouns, for example,

'mnozhestvo R' ("set R")

and 'mnogoobrazije R' ("manifold R").

Different N's in expressions of the type $N \overset{f}{\curvearrowright}$ (with the same f) can refer to each other as genus and species or can name objects for which the fact of their identity has been proven in the text. Using the axiom A_4 we can (very rarely) make an error. An error can occur in a case where the formula has the meaning specified once and for all independently from the text. So, without any previous definition of the meaning in an introductory construction or in a construction with the appositive link, a formula can <u>at once</u> be linked genitively to a noun.

This situation is not typical in mathematical texts. In this case we have a hieroglyphic word (cf. '&', '$' in common English) and not a freely chosen notation. Such a word must be stored in the dictionary (with the

specific meaning ascribed to it). For example, 'dx/dy' is 'derivative'.

Using Cl(f) in every case of occurrence of every formula, authors of mathematical texts would make the above procedures unnecessary. The problem of standardizing the language of scientific publications is not new, and in many cases some format of texts is prescribed.

The problem of choosing occurrences of genitive links in constructions of the type N ⌒ f from the set of all occurrences of such constructions in mathematical texts and, also, of choosing the only semantically relevant governor for a formula which has several formally equivalent ones is considered in (1). The general procedure for resolving ambiguities in surface syntactical analysis using broad context is proposed in (2).

R e f e r e n c e s:

1. S.A. Djulmisarjan, F.A. Dreizin, Z.T. Ter - Misakjants. Mathematical Formulae in Broad Context. Scientific & Technical Information, Series 2, No. 3, Moscow, 1971, p.p. 33-38.

2. F.A. Dreizin. A Computational Approach to the Choice of Analysis in the Case of Syntactic Ambiguity. Mechanical Translation and Applied Linguistics, No. 10, Moscow 1967 p.p. 3-20.