UDC 004.822/004.652.5

D. S. Sergeiev

A MODEL OF RELATION OBJECT FOR THE NATURAL LANGUAGE KNOWLEDGE BASE

This article focuses on forming a uniform object capable of representing a relation in a natural language knowledge base. The article describes both the functional requirements for such object and its resulting structure. Creation of this object allows to finalize the basic structure of the knowledge base and to implement it as a software product — both to test its capabilities and to use it as a foundation for following specialized extensions.

Keywords: natural language, knowledge base, relation, quantum of knowledge.

Introduction

Ever since the appearance of the first computers, unceasing attempts have been made to create an artificial intelligence – a more advanced version of the human intelligence, enhanced by the vast network of knowledge nodes and processing units. It was then, and remains now, a very ambitious goal which poses numerous challenges. One of them is programming a computer to understand and work with natural language information (NLI) – at least in the perfect synthetic form of structured texts.

A computer model of NLI is being researched in the department of Technical Cybernetics, FICT NTUU «KPI» as an attempt to resolve this problem. The proposed approach is based on several interrelated works by well-known scientists in fields of neurophysiology, psychology and linguistics. The key points of the research are the model of ISS (individual speech system) as a model of a human's speech processor and its integral parts – KB (knowledge base), where the NLI is stored, and LP (linguistic processor), which handles the two-way transformations of KB structured data and NL text data [1, 2].

The architecture of KB with lexeme-level elements has been developed in previous works. Now the main focus has changed to defining the links between lexemes as relations – separate objects that can be used to define the semantic load of the links, and, in this way, provide a ground for the next level of capabilities of the KB.

The goal of this article is to formally describe the model of relations between separate situations in KB of ISS. The objectives are to analyze the functions of relations in NL, to examine the particular qualities of the textual implementation of the relations, and, finally, to synthesize the relation object structure that can be used in the computer implementation of KB.

Theory

The concept of individual speech system (ISS), which is the ground for the whole model, is based on the research by L. Shcherba [3], and defines two parts of the human language processing core: the knowledge base (KB), where all known information is stored, and the linguistic processor (LP), which transforms and links sensory and symbolic information.

The main essential element of the KB is the basic semantic-syntactic structure (BSSS), which describes all parts of a single visual situation (*Obj*, *Mov*, *Attr*, *Attr*(*Attr*)). The structure of BSSS is based on S. Zeki's research of visual cortex [4]. This model of structure is confirmed by A. Gvozdev's observations on the process of learning speech by a child [5] and the idea of four main semantic parts of speech (*noun*, *verb*, *adjective* and *adverb*).

During several years of work on the KB structure, a model of single mono-predicate BSSS has been formed and tested [6, 7]. That allowed, to a certain extent, to map the structure of a visual situation (sensory data) through the textual representation to an artificial structure. Such structures, joined in a network, can now be used to determine the quantitative measures of the connections between concepts, but a more sophisticated structure of a relation object is still required to move it to the qualitative level.

Relations in KB

Limitations and capabilities

Up to this moment, our attention has been focused on creating a data model for storing and manipulating mono-predicate BSSSs – basic semantic-syntactic structures that represent a situation with a single subject and a single predicate. As it is based on the internal patterns of the real-world visual analyzer, the main part of this job was to find an adequate artificial model that can accurately represent the existing structure.

The problem of defining the relation object, while it exists in the same system of coordinates, is totally different, as this model cannot be based on any real-life object. On one hand, it is an obstacle, as no clear starting point exists to understand how it should function. On the other hand, though, it is an advantage, as any model that suits our needs at the moment given can be implemented.

Although we start with a *tabula rasa* relation model, the following prerequisites can be used to limit the space of selection.

The first one is that relations — just like BSSS — are originally stored in the human brains, and therefore they are bound to have quite simple biologically based structures that can be either transported through generations of genes or implemented via existing neural network capabilities.

The second one is that the number of inherent basic types of relations (such as time, space, casual, hierarchy etc.) is limited, and all the complexity comes from the sheer number of their implementations, which are theoretically unlimited.

The third one is that any possible limitation of the relation structure should only come from the corresponding biological structure itself, so there should be no strict logical rules and condition that are tied to arbitrary numbers, unless these rules derive from neurophysiology.

Basic model

Currently only 2 types of objects in the KB are known: a mono-predicate BSSS and a relation. The relation as a connecting element requires two BSSS's to work. S_o, the relation (R) can be represented as an entity that is linked to two BSSS objects (S).

$$R(S_1, S_2) \qquad (1)$$

Let us have a look at a primitive example of two BSSS's with a relation between them.

«The flowers smell strongly after a rain has passed»

Here we can clearly identify two situations: *«The flowers smell strongly»* and *«a rain has passed»*. Both are mono-predicate BSSS and both can range from one word to a whole complex structure — which is, however, still limited by 1 subject and 1 predicate.

Note that in this example the relation is obviously detected by the presence of the word *«after»*, which is not a part of any of the two BSSS’s. It is the main limitation of the relation: it cannot exist on its own, without links to BSSS's; neither can it be represented as one of the elements of the BSSS. Of course, as one of the major elements of the KB, the relation can have a very flexible and complex structure – consist of more than a single word, include grammatical structures, be implemented through inflection etc. The capabilities of this extension are covered in detail in the following sections.

Multiple relations

It is obvious that the relation model from (1) can only work with exactly two BSSS's. In real texts, however, the relations are often expanded either by homogeneous parts of sentence *(«the students and the teachers are in classrooms»)* or by chains of relations *(«the student of a faculty of a university»)*. The first problem can be solved by replacing the homogeneous parts by several similar constructs (*«the students are in the classrooms»* and *«the teachers are in the classrooms»*, but the chains of relations need to retain the integrity of links between the resulting decompositions. Therefore, the relations should be able to link not only exactly two BSSS's, but any arbitrary combinations of BSSS's and other relations, making them practically recursive.

Of course, when stored in KB, these relations have to be separated to preserve the contexts of different overlapping instances of relations, but the theoretical model should be capable of storing the chained relations directly.

Indication of relations in NL

As stated above, a human can identify the presence of a relation in a NL text without knowing the details of the said relation or even its type. A good example would be the «necessary» and «sufficient» conditions for logical statements, particularly theorems. These two types of conditions both belong to the group of causal (or logical) relations, but their meanings are distinctly different. Middle school students often confuse these relations and use them incorrectly. However, the students can easily point out the presence of relations between situations in a text, even if they misunderstand the types of the relations.

Since we assume that a NL text is only a reflection of a universal NLI structure, it is natural to expect that the relations in a NL text are already marked in one way or another. The exact nature and features of such markers is not a target of this article or even this research, as the LP is fully responsible for finding and identifying them. We need, however, to have a place for them in the relation object model.

The grammatical structure of relations is highly variable. It may contain joining structures (e.g. *«the student of a faculty»*), relation-specific keywords (e.g. *«day after day»*), grammar rules (*«if rain is falling, the ground is wet»*), or any permutation of the above. As far as it regards the KB, each of these rules can be expressed in a form of a grammatical rule, so only one additional field in the relation object is needed to contain it.

The new uniform relation structure looks like this:

$$(S_1) - (R) - [R_{\text{data}}] - (S_2) \quad (2)$$

The «Relation data» element is optional, as sometimes it may be redundant. In the majority of cases, though, it will be active, as even information about inflections is stored in it.

Open relations

As shown above, a relation requires two BSSS's to function. However, in NL texts the relation is sometimes left open – e. g. all signs of presence of a relation are evident, but only one structure is explicitly linked to it. Let us modify the example from the previous section and split it into two fragments: (1) *«When is the ground wet?»* and (2) *«When the rain is falling»*.

The second fragment contains a relational keyword *«when»* and a situation *«the rain is falling»*, though there is no indication of the other half of the relation inside it. It is obvious that the relation is linked to the first fragment – but any single fragment should be stored as a separate entity in the KB. Therefore, the model of the relation structure should be an optional element, as it may be absent in the text fragment to which the relation belongs. However, it is worth noting that from the semantical point of view the second structure is not optional, but semantics are on a higher level of the ISS model than the KB structure and should be dealt with separately.

Let us put down the new structure of the relation object, with the second structure being optional:

$$(S_1) - (R) - [R_{\text{data}}] - [S_2] \quad (3)$$

Semantical additions

Technically we could stop on the relation object presented in (3), as its structure is sufficient to store a relation. However, there is one more feature of the NL relation that is almost indistinguishable from the basic structural features. This feature is the type of relation. There are many possible ways for two situations to be linked, especially considering that a situation can consist of as little as one word – but there are also some clear widely-used patters that can be unified by group or purpose.

A hypothesis was presented in work [8] about existence of certain pool of types of relations, such as causal, spatial, temporal etc. According to this hypothesis, the relations are formed from the ISS and their types are identified not only by grammar (as in «*a student of a faculty*») or marker words («later», «in», «after»), but also by certain keywords – «kilometer» as a unit of space, «hour» as a unit of time etc.

This assumption opens up a possibility to label every relation by «type» based on its surroundings. For example, constructs «*2 kilometers to the city*» and «*2 hours to the city*» that follow a template «S1 *to* S2», are indistinguishable, but if we take into account the units of measure we can tell them apart and identify the types of the relations – temporal and spatial, accordingly. At the same time, «*2 overnight stops to the city*», while based on the same template, is tagged as a causal relation – as «*overnight stop*», whatever it is, is not a unit of measure and should be considered a separate situation. From the point of view of the KB, being a «unit of measure» is just an attribute of a word or a situation, which is applied by the «teacher» – a human user or an automated system.

One more semantic feature to note would be the direction of the relation. There are many symmetric relations that complement each other (*A near B = B near A*) or anti-symmetric relations that oppose each other (*A after B = B before A*). By adding the «direction» field to the type of relation we can drastically reduce the size of the corresponding database, as all the synonyms will be affected, too. As with the type of the direction itself, though, it is not stated as an inherit property of the relation – it is only added for ease of use and convenience and may be expanded, changed or even removed in the future.

Conclusions

Let us see how the complete structure of the relation object looks after all additions:

$$(S_1) - [R_{\text{direction, type}}] - [R_{\text{data}}] - (S_2) \qquad (4)$$

The relation is linked to 2 BSSS's, one of them being mandatory and another one optional. The additional data consists of grammatical data (details of how the relation IS implemented in a NL text), and semantic data (the type and direction of the relation).

As with all levels of the ISS model, the relations level is strictly additive: it does not modify existing KB structure and can be changed, removed or separated from the lower levels of KB without any additional costs or data lose.

This object should be able to store any relation with minimal or none modifications, provided that the LP detects and processes the required relation features from NL text.

Although there is a semantic part in the relation object, it is only added for convenience. Likewise, conflicting or ambiguous relations, if any, should not be resolved on this level.

In fact, the addition of the relation object finalizes the basic structure of KB. Higher levels, such as semantic level or source linking level, should be created and supported as separate extensions for the KB. The KB itself, even limited by the basic functionality of BSSS's and relations, is a complete product that can be tested, filled with knowledge and used as a framework for solving primitive tasks and creating extended systems that can solve corresponding specialized tasks.

References

1. Кисленко Ю. И. От мысли к знанию (нейрофизиологические основания) — монография / Ю. И. Кисленко – К.: «Український літопис», 2008 – 102 стр.
2. Kyslenko Y. Cognitive architecture of speech activity and modelling thereof / Y. Kyslenko, D. Sergeiev. // Biologically Inspired Cognitive Architectures. – 2015. – № 12. – С. 134–143.
3. Щерба Л.В. Языковая система и речевая деятельность / Л.В. Щерба. – Л.: Наука, 1974.
4. Zeki, S. (1992). A visual image in mind and brain: collection of papers. The World of Science, 11(12), 33–41.

5. Гвоздев А. Н. Формирование у детей грамматического строя русского языка / А. Н. Гвоздев. – Москва: АПН, 1949.

6. Сергеєв Д.С. (2014). Комп'ютерна оболонка природно-мовної бази знань (магістерська дисертація) / Д. Сергеєв. – Київ: НТУУ «КПІ», 2014.

7. Кисленко Ю. І. Структурний підхід до пошуку природно-мовної інформації / Ю. І. Кисленко, Д. С. Сергеєв. // Радіоелектроніка та інформатика. – 2015. – №3. – С. 45–49.

8. Кисленко Ю. І., Черевко О. С. Категорії часу та простору в інформаційних природно-мовних технологіях. — Автоматичні системи автоматичного управління.— 2011. — №13(38).