The ellipses in the geometric texts

Ksenia Naidenova
Military medical Academy
Saint-Petersburg, Russia
ksennaidd@gmail.com

Kurbatow Sergey
Scientific research centre
for electronic computer technology
Moscow, Russia
curbatow.serg@yandex.ru

***Abstract*—It is presented processing of ellipses occurring in texts of plane geometry tasks described in natural language. It is proposed a general approach to the processing of ellipses based on cognitive semantics. The ellipsis processing is based on the parallelism between syntactic structures and using geometric semantics. Some examples of the ellipsis processing with its limitations are described. The types of ellipses most commonly encountered in geometric tasks are highlighted. It is noted the ability to recognize ellipses and their resolution in the framework of cognitive semantics.**

***Keywords*—ellipse, a geometrical problem, cognitive semantics**

I. Introduction

The ambiguity of natural language caused by homonymy, has long been studied by computer linguistics. However, the ambiguity associated with the omission of a thinkable language unit (ellipsis) in the text has been actively analyzed by automatic processing natural language relatively recently [1], [2]. Although, in the theoretical linguistics, the ellipticity got enough coverage [3], [4], the restoration of ellipses in the systems of syntactic text analysis is clearly developed not enough. This is largely due, first, to the fact that eliminating the ellipticity is of a subordinate nature with respect to the actual syntactic analysis and, secondly, to the complexity of recovering ellipses. The complexity is explained by the potential need to take into account a number of contexts: current sentence, adjacent sentences, already established syntactic relations and, finally, semantics of text.

This work is divided into two parts. In the first part, it is described how to handle the ellipticity in a specific holistic system of solving plane geometry tasks in natural language description. The second part proposes a general approach to the processing of ellipses based on cognitive semantics, aimed at a much broader range of applications (both in subject areas and in different natural languages).

A. Treatment of ellipses in a system for solving geometrical tasks automatically

The general scheme of an integrated system to solve geometrical problems automatically in natural language is given in [7]. The system includes a linguistic translator, an ontology and a graphical interface for output result (drawing of NL-explanation of the solution process). An expansion system provides the interpretation of more complex tasks (to identify their logical structure) associated with language processing in the limits required by a number of modifications of the parser. This section describes an extension of the system to correctly interpret elliptical (incomplete) sentences. The correct interpretation of the ellipses is based on the parallelism between the syntactic structures (the ellipticity) and knowledge about the general concepts included in the structural schemes of ontology. Therefore, the interpretation essentially uses both levels of linguistic processing - syntactic and semantic ones.

The language translator creates a syntactic structure and determines that some of its elements violate the language rules. For example, there is no noun for adjective, pretext is at the end of sentence, number does not have a mandatory measuring unit, and so on. The basic criteria for determining ellipticity are studied by linguists [6]. Based on these criteria, recorded in the ontology, the translator identifies the fragments of syntactic tree that admittedly contain ellipticity. Next, with the use of algorithms described in short below in 1.2, the identified ellipses are restored.

Specifically, in sentence «the radius of the first circle is set to 12 cm, and the second 10», the elements «second» and «10» define the ellipticity. As a result, two syntax structures must be formed:

- The radius of the first circle is 12 cm.
- The radius of the second circle is 10 cm.

These structures are further processed by the system mechanisms of paraphrasing to obtain an ontological representation of sentence in formal terms of the subject area [7].

The concept "‘paraphrasing"’ has been proposed by the well-known Russian linguist Apresyan in [10]. In our system, we use an adaptive variant of this concept. The conception of paraphrasing assumes that any class of sentences corresponding to one and only one sense can be reduced to the simplest or canonical phrase composed only of the lexemes expressing most clearly the based concepts of the sentences. Thus "‘paraphrasing"’ is based on the following proposition in [10]: «One of the fundamental properties of human languages consists in the fact that if there are several synonyms, in the broad sense, to express some concept, then only one of them turns out to be privileged, canonical, or prototypical for expressing this concept.» [10].

In particular, such canonical concepts in the plane geometry are, for example, «point», «line», «plane» and «belong», «to lie between» and «to be congruent».

In the « Space shipyard » domain [8], the canonical concepts are «tank», «adapter», «reinforcement», «dock», «point of joining» and some others.

Thus, the rules of paraphrasing provide only one canonical form for a group of sentences having the same sense. For example, sentences "a point located on the straight line", "the straight line passing through a point", "a point belonging to the line", "a point lying on the line segment", etc. are reduced to the following canonical phrase, namely, "point belongs to straight line".

This canonical phrase is displayed in the ontological representation in the form of the following triplet « point lies line ». It is noted to emphasize that the members of the triplet (objects and relations between them) are not dependent on a particular language. Therefore, the corresponding rule of paraphrasing contains, in its left part, the objects and relations depending on the language, but, in its right part, the objects and relation are invariant with respect to the different languages.

The rules of paraphrasing are divided into two classes, the first one consists of rules in which both parts of them are some generalized syntactic structures; the second one consists of rules having the canonical descriptions in their left parts and the semantic descriptions in their right part. The second class rules can be used for transforming the ontological structures into the corresponding natural language texts. It is reasonable to apply the rules of the first class for equivalent synonymic transformations of the synthesized structures to retrieve texts in the most appropriate manner in a considered application domain.

B. Algorithm for resolving ellipticity

B. The processing algorithm is based on the ontology knowledge reflecting the semantic hierarchy of word forms in the syntactic structure and rules of natural language. To a first approximation, the algorithm can be described as follows:

- to segment syntactic structure into two segments: a complete one without ellipticity and the other containing an ellipticity (generally, there are noun groups);
- in the elliptical structure, identify elements that can be usefully matched with full syntactic structure elements to be used for resolution of ellipticity;
- in the full syntactic structure, identify candidates to replace with the elements found in the previous step;
- perform a replacement and get a complete syntactic structure.

In the example from subsection A «first» is replaced by «second» and «12» by «10» because they correspond to the same concepts of ontology. Here we have different objects and the same type of attribute (length). In sentence «the perimeter of the triangle is 37 cm and area – 20 cm», we have the same object and the different types of attributes. This seemingly simple algorithm allows to successfully recover not only geometrical ellipses, but several others, described, for example, in [2]: In sentence «twenty years of this dance form the age, forty – the history», «twenty» is replaced by «forty» and «age» is replaced by «history».

Naturally, the ontology should contain the concept «time-interval» binding hierarchically «age» and «history». It is obvious that the standard tools for editing ontology ensure that the algorithm is correctly carried out without reprogramming. More examples: «He went to the pharmacy and his brother – to the post office». «He» is substituted by «his brother» (It is a noun group: whose brother? him). Here the noun group could be: «its half-brother on the maternal line». The mapping in the ontology is formally performed for the roots of words in syntactic tree of noun groups. It is for this reason that the structure for sentence «Triangle ABC is inside the circle and the square – outside» is correctly restored.

C. Limitations

Of course, many cases of ellipticity can not be processed by the above algorithm. Example: «There are seven circles. Radius of one 5 cm, two other – 3, and the others – 10 cm.» We have multiple ellipticity in this example. A similar example from [2]: «Anemones discard tentacles, crayfishes – claws, lizards – tail». In some cases, there is an ambiguity at the comparison level. At the logic level, two options were analyzed:

1) continuation of work and eliminating ambiguity at the stage of semantic processing (canonical syntactic structures);
2) enter in to the ontology not only knowledge about the hierarchy of concepts, but the rules of preferences when choosing a candidate for replacement (substitution).

A decision on whether to choose one of the options or a combination of them is the subject of further study. In any case, the focus is on the universality of knowledge-based algorithm. Note that the algorithm described was tested not only for geometric texts but also for a number of others, in particular in the text of so-called «space shipyard»[8]. The ontology fragment, which describes the concepts of this area, specifies the hierarchy (visualization object) isA (tank, adapter, armature, solar panel, energy block, and so on) isA (specific types of objects: Tank-b, Truss-c, etc.). According to this hierarchy, sentences of the type «diameter of the first tank equals 1.7 cm and the second 5.8» or «the length of the solar panel is 15 cm and the weight 40 grams» is successfully restored.

It should be noted that the question of a clear ellipticity criteria and methods for restoring the full structure of sentences has not been fully resolved within the framework of a generally accepted linguistic theory. Further, the following cognitive approach to the resolution of ellipses provides a significant extension of the possibilities described above and focuses on self-learning ontologies for the treatment of a wide spectrum of incomplete sentences.

II. Ellipses processing based on cognitive semantics

1) General information: The cognitive approach is opposed to traditional formal one, based on Tarski and Hilbert's ideas, playing a major role in the computer paradigm of most AI research. Two main differences of the cognitive approach from the traditional formal one are to solve categorization and semantics problems. Categorization is a problem of creating concepts (categories) and structuring the conceptual human system.

Resolving ellipses in natural language texts remains one of the most difficult and unsolved tasks in linguistics, despite the abundance of proposed methods based on semantic-parsing of sentences. The syntax reveals the structure of ellipse and the similar part of the sentence without the ellipse, and the semantics deal with word values. However, as the example from (Umberto Eco. The role of the reader. Exploration in the semiotics of texts. – Moskva: Publishing House ACT: CORPUS, 2016 (page. 62)) shows, resolving ellipses is based on the understanding of context (text theme), the meanings of words and collocations: «Charles makes love with his wife twice a week. So does John».

2) Ellipsis classification in geometrical tasks: To study the typology of ellipses in geometric tasks, we used a body of texts that is listed in [9]. From a variety of sentences, we have selected several types of ellipses: ellipses using the sign «-» (ellipses with missing a predicate, ellipses with missing a verb), ellipses without using the sign «-» (ellipses with a skipped noun, with a skipped pronoun, with a skipped predicate. The characteristics of these types are analyzed, not necessarily related to the subject area (geometry) and the language (Russian).

A detailed discussion of the selected ellipses in terms of their properties and structure is given in [9]. Resolving complex cases of ellipticity requires understanding the context of geometrical tasks. To account for context, we enter the concepts of cognitive models of geometric object and action with object. The model of object includes the properties of the object, the actions as a result of which this object can be created, the actions that object can perform itself, and in which it can participate; the object model also includes the elements of which object is composed, and the elements of which it is a part. Cognitive schema will show how the object is formed and its position in space in respect to other objects. The model of action, naturally, includes objects involved in this action and its result as a geometrical configuration of objects. And all the parts of the cognitive models will be associated with the words that appear in the text body in question.

Within the proposed model, text analysis becomes cognitive-driven, and the parser plays a subordinate role in the process. In processing ellipses resolution based on cognitive models, it is possible to synthesize text that describes the geometric situation and compare the text generated with the text to be analyzed. Ontology contains theoretical knowledge in the area of plane geometry and knowledge of methods of solving plane geometry tasks of various types (computational, construction, challenge for proof). Ontology takes the burden of solving the problem and creating the drawing accompanying solution (visualization of solution). The Cognitive Analyzer of text runs incrementally and transmits the converted and meaningful text to the ontology in the language required by the ontological block.

A. The structure of cognitive models of objects and actions

The cognitive structures correspond to the semantic structures of the situations described in text. They should be aligned with the narrative structures of sentences. A word can have multiple values, but only one meaning, at least in mathematical texts. Ellipsis (omitting words, economy of text) is possible because the preceding text determines unambiguously (uniquely) the meaning of each word and situation, and the meaning remains unchanged. In cognitive model of object, the following relationships are important:

- object can perform some actions;
- object can be subjected to action of other objects;
- object can have spatial and temporal relationships (earlier, later, already built, already given) with other objects;
- object can be composed of some other objects;
- object can be a part of some other object (objects);
- object has properties, some of which (call them actant ones) are related to the actions that the object commits (intersects-intersecting, lies - lying) or the actions that are committed over it (has been given - given, has been formed – formed, cut off, embedded). Thus, the actant properties of objects are directly displayed in the morphological forms of the words describing these properties;
- the relationships between the properties of one geometrical object, between object and its parts are realized through implications: if "‘center"’, then "‘a circle"’; if "‘radius"’, then a "‘circle"’; if "‘circle"’, then "‘circumscribed about or inscribed in"’; if "‘inscribed in"’, then "‘in an object"’; if "‘bisector"’, then "‘bisector of an angle"’; if "‘bisector of angle"’, then "‘the vertex of angle from which it originates"’; if "‘bisector"’, then "‘the angle from which it comes is divided in half"’; if "‘bisector"’, then "‘it is the axis of symmetry of angle divided in half by this bisector"’.

With cognitive patterns, we associate such phenomenon as a cognitive wait for appearing certain words and narrative constructions in the text. All cognitive models can be explicitly defined based on geometric semantics, and are associated with speech parts and typical collocations with their grammatical categories at the sentence level. The creation of cognitive models of objects and actions for plane geometric tasks in the proposed approach is performed in a step-by-step mode by the use of a given text corpus. An example of creating a cognitive model "Bisector" across the text of many tasks is shown in [9].

It is a problem of considerable interest to apply plausible reasoning for the resolution of ellipses, including analogy, generalizations, specialization, use of implications, forming hypothesis and many others. Example 2 (Section 1.3) “Anemones discard tentacles, crayfishes – claws, lizards – tail” can be resolved by analogy. Anemones, crawfishes, lizards are living things, tentacles, claws, tale are parts of their bodies, then crayfishes, claws and lizards, tail can be connected by the action "discard".

III. Conclusion

The processing of ellipses is given in a specific system of plane geometry tasks with a natural language description. The ellipse resolution is based on using in parallel the syntax structures of sentences and the geometry semantics. A broader approach to ellipses processing based on cognitive semantics has been proposed. The approach gives a classification of ellipses (across a geometric text corpus) and introduces the cognitive models of geometry objects and actions. The model proposed allows to view the structure of automated analysis of geometric texts as a cognitively controlled parsing. Further research is envisaged in two directions:

1) enhancing the capacity of the existing system;
2) algorithmization and software implementation of the cognitive approach.

Acknowledgment

The work is executed at financial support of RFBR, project No 15-07-03847, "Interpretation of the technically-oriented natural language texts using application ontologies" .

References

[1] Malkovsky M. et al. Restoring ellipse as a task of automatic text processing. Program products and systems, No. 3 (107), 2014, http://www.sworld.com.ua/konfer35/365.pdf

[2] Kobzareva T., Epifanov M., Lakhuti D. Finding the antecedent of the ellipsis of a fragment containing predicate (automatic analysis of Russian sentence). The Proceedings of 15th National Conference on Artificial Intelligence wth International Participation CAI-2016, Vol. 2, pp. 56-62.

[3] Adametc P. Semanticheskaya interpretatsiya «znachimykh nulei» v russkikh predlozheniyakh [Semantic interpretation of "significant zeros" in Russian sentences]. Proceedings of papers in honor of Dean C. Worth, for his 65 anniversary. M.: East Literature, Russian Academy of Sciences. 1995. P. 9-18. (in Rus)

[4] Vardul I. To the question of the ellipse phenomenon //Invariant syntactical values and structure of sentence. M.: Nauka, 1969. P. 59-70.

[5] Minjajlov V. Detection of elliptical sentences by means of classification algorithms. In: Problems of Physics, Mathematics and Technology, No. 3 (28), 2016.

[6] Larkina, A. Ellipsis in the modern French language, The thesis of a PhD in Philological Sciences (specialization 10.02.05). 2009.

[7] Kurbatov S., Vorobyev A. Ontological Solver of Geometry Problems in natural language description. // Proceedings of 15th National Conference on Artificial Intelligence with International Participation (CAI-2016), Vol. 1, P. 56-63.

[8] Lobzin A., Khakhalin G., Kurbatov S., Litvinovich A. Integration based on natural language and image ontology in the system Text-To-Picture. The Proceedings of 8th Scientific-Practical Conference «Integrated Models and Soft Computing in Artificial Intelligence», Phizmatlit, 2015, 296-305.

[9] Processing of ellipses based on cognitive semantics, http://www.eia–dostup.ru/cognitiv.htm

[10] Yu.D. Apresyan, I.M.Boguslavskii, L.L.Iomdin i dr. Lingvisticheskoe obespechenie sistemy ETAP-2 [Lingware of the ETAP-2 system] M.: Nauka, 1989.

[11] Eco Umberto. The role of the reader. Exploration in the semiotics of texts. – Moskva: Publishing House ACT: CORPUS, 2016 (page. 62))

ЭЛЛИПСИСЫ В ГЕОМЕТРИЧЕСКИХ ТЕКСТАХ

Найденова К. А.
Военно-медицинская академия
Санкт-Петербург, Россия
Курбатов С. С.
НИЦЭВТ ОАО - АО «НИЦЭВТ»
Москва, Россия

Аннотация. В работе описана обработка эллипсисов в конкретной системе решения планиметрических задач по описанию на естественном языке и предложен общий подход к обработке эллипсисов на основе когнитивной семантики. Разрешение эллипсиса базируется на параллелизме синтаксических структур и использовании семантики геометрии. Даны примеры обработки и описаны ограничения. Выделены типы эллипсисов, наиболее часто встречающихся в геометрических задачах. Предложен подход к распознаванию эллипсисов и их разрешения в аспекте когнитивной семантики.

Ключевые слова: эллипсис, геометрические задачи, когнитивная семантика.