A TEXT GENERATION SYSTEM FOR EXPLAINING CONCEPTS IN GEOMETRY

Ruslan Mitkov

Institute of mathematics
P.O. Box 373, Sofia
Bulgaria

Introduction

The project note presents the computer program GECO (GEometry COnsultor), which generates explanations (descriptions) of geometrical concepts in Bulgarian. The program replies to questions such as "What is X", "Compare X and Y", where X and Y are concepts in geometry. The idea to design GECO was initially inspired by [McKeown,1985], which gives information on missiles. However, unlike TEXT, our system is able to give two types of descriptions : detailed and brief. For instance, if we ask "What is a triangle", GECO gives a detailed description, including definition, the characteristic and non-characteristic features of the triangle, but if we ask once again within a dialogue the same question, the program reminds us shortly the definition of the concept triangle. This is done by means of a special stack, which keeps a formalized dialogue history. As in normal human dialogue, the systems replies to a repeated question in past tense (this is specific at least of Bulgarian). Here is a "recorded" dialogue between user and GECO:

User: What is a triangle?

GECO: The triangle is a convex straightlinear plane geometrical figure, which possesses 3 sides. According to its sides the triangle can be equilateral, isosceles and scalene and according to its angles - right-angled, acute-angled and obtuse-angled.

....

User: What is a triangle?

GECO: The triangle was a convex, straightlinear and plane geometrical figure, which has three sides.

In our model of natural language generation, we assume, that the task of language generation can be divided into two stages: a conceptual and a linguistic stage. We agree with Danlos' claim [Danlos, 1985], that the conceptual and linguistic stage cannot be completely independent. That is why not all decisions in the conceptual stage of GECO are to be fixed invariably.

In [Roesner, 1986] is given a typology of explanations. According to this typology, the explanations, generated by our system are "informative explanations, of static kind" (this type includes the explanations of concepts, what is our case). Roesner proposes, that this type of explanations should possibly include definitions, short descriptive information, indication, where more information can be found, relations between concepts etc. Originally, GECO was intuitively so designed, that it practically met these formal demands.

Sublanguage and discourse

The system has to do with the sublanguage of school geometry (taught in bulgarian schools). This sublanguage uses a comparatively restricted lexicon and not too great variety of syntactic constructions. Its texts (what is to be read in the textbooks), however , are often overloaded with very complicated complex sentences. Good school geometry texts are presented in balanced way by simple and complex sentences.

Studying discourse pecualiarities of school geometry instructional texts helped us to design discourse rules, made use by the conceptual module, when ordering the content within a text. Generation of text requires the ability how to organize individual sentences. A reasonable writer does not randomly order the sentences in his text, but rather plans an overall framework or outline, from which the individual sentences are produced. Characteristic of the description of a geometrical concept is the introduction of its superordinate, its constituents and providing some additional information to it (e.g. varieties). In this way the description of the geometrical concept "quadrilateral" possibly includes its superordinate (polygon), its constituents (4 straightlinear sides, which build up a convex figure and lie in one plane, i.d. it has 4 sides and is convex, straightlinear and plane) and its varieties (parallelogram, rectangular etc.).

Semantic knowledge representation model

The semantic knowledge representation model used in the system and proposed by the autor is an extension of the model of Tiemann and Markle [Tiemann and Markle, 1978] for concept semantic knowledge representation. The proposed model describes each concept as a set of critical and variable attributes. The concept introduces a class of things, objects, events, ideas or relations, so that each member of this class possesses the same label. On the other hand it is possible that all the members of a class differ in one way or another and nevertheless are classified together. The characteristic features, possessed by all the members in a class are called critical attributes. Variable attributes are defined as characteristic features, which might differ within class members.

Consider the concept "triangle". Our semantic knowledge representation model will describe it formally as follows:

Triangle (geometrical figure / plane, convex, straightlinear, three sides / acute-angled, right-angled, obtuse-angled; equilateral, isosceles, scalene/O).

The formalism of functional descriptions

Different formalisms require different approaches, whose variety may be sometimes (especially in implementation) problematic. We have adopted in our model and system an extended functional description (FD) formalism, developed by Rousselot [Rousselot, 1985]. This formalism enables the representation of all types of knowledge. A FD represents a list of attribute-value pairs. Rousselot's formalism is a very extended form of the functional grammars [Kay, 1985]. Within the notation of FD we represent in the domain knowledge base the geometrical concepts (using the above concept semantic representation approach) and the relations among them . We represent also as FDs the grammar rules in the linguistic knowledge base.

The role of logical emphasis

Different text generation systems make use of different syntax selection approaches. The phenomena of focus is widely used in text generation [Derr and McKeown, 1984], [McKeown, 1985]. In brief, if the focus is on the protagonist of the sentence an active construction is chosen, and if the focus is on the goal - a passive one. Thus if the system TEXT [McKeown,1985] should express the fact, that a triangle and circle touch each other and if the focus is on the protagonist (say triangle), the system would generate "The triangle touches the circle", but if the focus is on the goal (say circle), the folllowiing sentence would be produced "The triangle is touched by the circle". Danlos [Danlos,1985] makes use of "discourse grammar", that specifies the syntax of each sentence.

In our work we have elaborated for the first time on the logical emphasis approach for the sentence syntax selection. The logical emphasis in free order Bulgarian is laid on the last word of the sentence. It plays a leading role in choosing the word order in a sentence: the fact that the triangle has three sides may be verbalized as "the triangle has three sides" if the logical emphasis is on "sides" or as "three sides has the triangle" provided the logical emphasis is on "triangle".

Production of text

The grammar used to produce the preliminary surface structure of the generated sentences is the functional unification grammar [Kay, 1985]. If a simple sentence grammar pattern (subject verb object) with value "protagonist" for the funcional role (attribute) "logical emphasis" is unified with the FD from the domain knowledge base, describing the relation, that each triangle has three sides, the following sentence will be generated:

The triangle possesses three sides.

while the value "goal" for the attribute "logical emphasis" yields the result:

Three sides has the triangle.

However, the sentence generated by unifying the functional unification grammar with the input and representing a fact about or related to a concept is not the final sentence the system offers to its users. Although the resulting sentences of the previous examples sound quite reasonably (in Bulgarian), the system would not be able to impress always its users if it accepts the sentence as final. The problem is that each explanation of a concept is not a single sentence, but discourse. To illustrate our position, assume that the system has to give a detailed (initial) description of the concept "triangle" (such description is actually given by our system, see in the dialogue first answer). After consecutive unifications of the grammar rules with the relevant inputs , the system would generate in the best case the following text:

> The triangle is a geometrical figure. The triangle is straightlinear. The triangle is plane (plane as adjective). The triangle is convex. The triangle has three sides. The triangle can be isosceles, equilateral and scalene according to its sides. The triangle can be right-angled, acute-angled and obtuse-angles according to its angles.

This is a clumsy text, that no reasonable man would write. The three main linguistic operations (part of the system's linguistic knowledge) that will process this priliminary text are coordination, pronominalization and ellipsis. Note that coordination will work on the first four sentences, pronominalization - on the fifth sentence and ellipsis - on the sentence obtained from coordination of the last two sentences (before these three operations a rhetoric rule will have operated, which says, that "according to" sounds better at the beginning of the sentence and is not subject to deletion during coordination). The processed text will be:

> The triangle is a straightlinear. plane and convex geometrical figure. It has three sides. The triangle can be isosceles, equilateral and scalene according to its sides. The triangle can be right-angled, acute-angled and obtuse-angled according to its angles. According to its sides the triangle can be equilateral, isosceles and scalene and according to its angles - right-angled, acute-angled and obtuse-angled.

There are also further linguistic decisions to be made: should the sentence be in active or in passive voice, should two or more simple sentences be combined into a single complex one (in the last sample text the first two sentences can be combined into a complex one), how subordinate clauses should be handled (we have developed several procedures to treat the production and connection of subordinate clauses), should gerundium be used etc. In text generation systems such decisions are made on the basis of linguistic phenomena such as focus, logical emphasis (in our case), causality etc. and are not to be discussed in the present paper.

Grammatical accordance

Since Bulgarian is a highly inflective language (inflection affecting not only nouns, but also adjectives, numerals, pronouns etc.), we have developed additional algorithms for grammatical accordance. We have developed an algorithm, which determines automatically the gender of the Bulgarian nouns (consisting of 254 steps). Another algorithm gives the definite article (in Bulgarian as a inflection; there exist various definite article inflections) of each noun and works parallelly to the first one. However, if a Bulgarian noun is in its definite article form, it is impossible to determine algorithmically its gender. Therefore we have developed additional algorithms for transforming definite article form of nouns into indefinite article (normal) forms. Moreover, the adjectives, numerals, the demonstrative and personal pronouns in Bulgarian accord with the nouns. Consequently we have developed and implemented algorithms for determining the indefinite article form of adjectives (numerals, pronouns) and from it the gender form and definite article form of adjectives (numerals, pronouns).

Implementation

GECO is a program, designed for instructional and experimental purposes. Its most part has been already programmed on IBM PC/XT/AT (in Ksi Prolog). The FDs are described within Prolog Definite Clause Grammars (DFG) notation. Thus we have implemented a surface generation based on both the DFG formalism and the formalism of FDs. This idea we have adopted from Derr and McKeown [Derr and McKeown, 1984]. The result is a generator with the best features of both grammars: simplificatiion of input by using functional information and efficiency of execution from Prolog.

References

Danlos, Laurence - Generation automatique de textes en langue naturelle, Masson, Paris, 1985

Derr, Marcia and McKeown Kathleen - Using focus to generate complex and simple sentences, COLING, 1984

Kay, Martin - Parsing in functional unification grammar. In Zwickky et al (eds): Natural language parsing, Cambridge, 1985

McKeown, Kathleen - Text generation: using discourse strategies and focus contstraints to generate natural language text. Cambridge university press, Cambridge, 1985

Mitkov, Ruslan - A knowledge representation model and its applications, Models of meaning, Varna, 1988

Roesner, Hannelore - Generierung von Erklaerungen aus formalen Wissensrepraesentation. Bericht No.3, Verbundvorhaben WISBER, 1986

Rousselot, Francois - Un systeme comprenant des textes en utilisant un formalisme unique. T.A. Informations, No.2, 1985

Tiemann P., Markle S - Analyzing instructional content: A guide to instruction and evaluation. Champaign, Illinois: Stipes publishing company, 1978