Anna Armeeva, postgraduate student (anna@bitsoft.ru)
Dr. Irina Kobozeva (logos@logos.msu.su)

Department of theoretical and applied linguistics
Philological faculty
Moscow State University

ALGORITHM FOR AUTOMATIC DESCRIPTION OF A PICTURE

Abstract

This paper describes a model for a computer program that must generate a text according to a picture. The picture belongs to one of three types - interior, landscape or still life. The program takes into account the cognitive abilities of a human being connected with the estimation of objects which is a necessary step for a verbal description of a picture. The program starts by writing a list of objects represented by their names and coordinates. Each object receives its weight. The weight of an object is calculated as a sum of its prototypical and variable weights. Then the program selects a mode of describing one of the picture types. Prototypical weight is built on the basis of such prototypical properties of the object having this name as its typical size and typical degree of stability/instability. The third property is the type itself of an object (slots of a frame, animated objects, etc.). These properties influence the order of appearance of objects in a text, the choice of this object as Figure or Ground (=Landmark, reference object). The variable weights depend on the coordinates of an object. The variable weight of an object is determined by the actual properties it has on a given picture: isolatedness, remoteness, partial representation, size relative to the other objects, singularity/plurality. The next stage is connected with taking into account the location of the objects and the planning of description. The last stage is text production.

This paper describes a model for a computer program that must generate a text according to a picture. The picture belongs to one of three types - interior, landscape or still life. The program takes into account the cognitive abilities of a human being connected with the estimation of objects which is a necessary step for a verbal description of a picture. In this program the objects are compared in their size, stability. The third property is the type itself of an object (slots of a frame, animated objects, etc.). All these properties influence the salience of an object. This salient object appears in the beginning of the text or a text fragment, is chosen as the Figure. There are prototypical weights contained in the memory of a human being in a “knowledge base” and variable weights connected with the properties of a real picture in the program. These weights permit to model the possibility of text generalization on the basis of visual perception. Thus two cognitive processes - verbalization and visual

perception are brought into correlation. One of the essential cognitive principles uniting these processes is the perceptual difference between the Figure and the Ground. This difference is subsumed under the notion "salience" as it is used in cognitive linguistics. The base of the salience is the unequalty of the parts of a picture and a text. In our program salience is modeled with the weights forming the hierarchy of the describing objects. It is the salience on the level of the knowledge organization. As for the text organization, the salience of an object is reflected by the order of its appearance in a text and its semantic role in a locative sentence (Ground vs. Figure).

We do not analyze the stage of objects' image recognition. It is assumed that all the objects are already recognized and every object has a name, denoting its category. The program starts by writing a list with names and coordinates of the objects. The objects have two x-coordinates, two y-coordinates and one z-coordinate. The objects' shape is a flat rectangle without thickness.

$N_1[(x_1\ y_1)\ (x_2\ y_1)\ (x_1\ y_2)\ (x_2\ y_2)\ (z_1)]$ [additional information],
$N_2[(x_1\ y_1)\ (x_2\ y_1)\ (x_1\ y_2)\ (x_2\ y_2)\ (z_1)]$ [additional information],
, … ,
$N_k[(x_1\ y_1)\ (x_2\ y_1)\ (x_1\ y_2)\ (x_2\ y_2)\ (z_1)]$ [additional information],

where N_i - a name of an object, $[(x_1\ y_1)\ (x_2\ y_1)\ (x_1\ y_2)\ (x_2\ y_2)\ (z_1)]$ - coordinates of the objects, $(x_1\ y_1)$ - a left lower point, $(x_2\ y_2)$ - a right lower point, $(x_1\ y_2)$ - a left upper point, $(x_2\ y_2)$ - a right upper point, (z_1) - a z-coordinate.

The names of objects are contained in a base and belong to one of three groups - "Interior", "Landscape" or "Still life". If the list of objects contains objects belonging to the group "Landscape", the mode of description "Landscape" will be selected. If all the objects not belonging to the lower level (it means they have the smallest y-coordinates) belong to the group "Still life", the mode "Still life" will be selected. If all the objects belong to the group "Interior", the mode "Interior" will be selected

Each name receives a prototypical weight that is built on the basis of some prototypical properties of the object having this name. These properties are typical size and typical degree of stability/instability. The third property is belonging to one of the types.

An object receives the mark 0, if it is stable. If an object is unstable, it receives -1. The objects are considered as unstable, if they are animate or can be moved easily (statuette, cup, plate etc.).

The typical size of an object depends on belonging of this object to one of the group. If the object "person" belongs both to the group "Interior" and the group "Landscape", it receives different marks in these groups. It depends on the context. In the interior the person is more important than in the

landscape. And a plate belonging to the group “Still life” differs from a plate in “Interior”.

In the group “Interior” an object receives following marks: 1 (if it is smaller than a person), 2 (if it is equal to a person in size) or 3 (it is larger than a person). In the group “Landscape” an object receives: 1 (it is point-like and equal to a person in size), 2 (it is point-like and larger than a person), 3 (it is point-like and much larger than a person), 4 (it is linear and much larger than a person) or 5 (it has a large area and is much larger than a person). In the group “Landscape” an object receives: 2 (small), 3 (middle), 4 (large).

The third property is the type of an object. There are following types: slots of a frame, animated objects, parts of objects, supports (tables, chairs, divans, etc.), covering objects (statues, tableware, etc.). The belonging of an object to one of the types influences the order of appearance of objects in a text, the choice of this object as the Figure or the Ground in such sentences as “the bike is near the house”.

An object receives following marks as belonging to one of the types: in the group “Interior” all animated, covering objects, parts of objects receive the mark -1. The rest of the objects receive 0. In the group “Landscape” all animated objects receive the mark -1. The rest of the objects receive 0. In the group “Still life” all covering objects receive 1, the rest - 0.

All the marks are summed up, this sum is the resulting prototypical weight.

Group “Interior”

	type	size	stability	sum
wardrobe	0	3	0	3
divan	0	2	0	2
table	0	2	0	2
support	0	2	0	2
chair	0	1	0	1
person	-1	2	-1	0
door	0	3	0	3
window	0	2	0	2
statue	-1	2	0	1
statuette	-1	1	-1	-1
stove	0	3	0	3
picture	0	1	0	1
mirror	0	1	0	1
chandelier	0	1	0	1

Group “Landscape”

type	size	stability	sum

tree	0	2	0	2
building	0	3	0	3
bush	0	1	0	1
river	0	4	0	4
field	0	5	0	5
lake	0	5	0	5
road	0	4	0	4
person	-1	1	-1	-1

Group “Still life”

	type	size	stability	sum
plate	+1	2	-1	2
table	+1	4	0	5
wineglass	+1	2	-1	2
vase	+1	3	-1	3
grapes	0	2	-1	1
tray	+1	4	0	5

The prototypical weights reflect some properties of an object-prototype. These properties influence the order of appearance of objects in a text, the choice of this object as the Figure or the Ground. The name of an object may be connected with some variable weights that are summed up with the prototypical weight. The variable weights depend on the coordinates of an object.

The list of the variable weights:

isolatedness 0 (standard) -1 (isolated object)

size 0 (small object) +1 (great object)

singularity/plurality 0 (no plurality) +1 (plurality)

remoteness 0 (foreground) -1 (intermediate space) -2 (background)

partial representation 0 (complete representation) -1 (partial representation)

The variable weights are characterized in following way:

isolatedness - the object is isolated if its length is less than the distance between the adjacent objects.

size - The adjacent objects having the equal prototypical weight are compared in area. The greater object receives +1, the smaller - 0. We compare the objects on one vertical line not belonging to the most lower level and the objects on one horizontal line belonging to the most lower level.

singularity/plurality - two or more objects with equal names and equal sizes belonging to one level and being close build a set. It means that they receive +1 in plurality. They will be described together (ex. “I see books” instead “I see a book”). If sizes of the objects with equal names are different,

the objects do not build a set. In this case one of them becomes the Figure, the other - the Ground. The adjectives “great” and “small” are used.

“One horizontal line” is defined in following way: the objects are on one horizontal line, if they cross the vertical drawing from the upper surface of the highest object or if they are below than this vertical. The z-coordinate of all the objects is equal.

“One vertical line” is defined analogous: the objects are on one vertical line, if they are within the limits of the verticals drawing from the lateral sides of the object belonging to the lower level. The z-coordinate of all the objects is equal.

Remoteness - realized only for intermediate space and background.

Partial representation - some objects can be represented only partially. The information about the partial representation is presented with the coordinates of the objects in the field “additional information”.

The next stage is connected with taking into account the location of the objects and the planning of description. The object with the maximal weight is found, it is the Main Object. No more than two objects are taken in all horizontal and vertical directions. If there are more than two objects, the next objects are not described with the Main Object. If the length of each object to the right and to the left of the Main Object is less than the distance between the Main Object and these objects, they do not describe together with the Main Object. The object with the maximal weight is selected from them, etc.

If the objects with equal names not belonging to one set are described together, the adjective “another” is put in. If there are two pairs of objects with equal names, the words “one more” are put in before the second object of the second pair.

Trajectory of description. There are many types of description trajectories [Porzel 1996, Kobozeva 1995]. Our description strategy is a point-by-point strategy that is anchored at some objects. And this anchorage is connected with the properties of these objects. Firstly the group of the Main Object is described, then the following object with the maximal weight is selected, its group is described, etc. If there are several objects with maximal weight, the description moves from left to right.

Description in groups. The general scheme is following: firstly the object with the maximal weight is described, then - the upper objects, then - the front objects, then - the back objects, then - the left objects and the right objects. The latter four groups can be enlarged with the description of upper objects. Each group can contain no more than two objects. If some front or back objects are at the same distance from the Main Object, we describe firstly the object with the greater weight. If there are more than one such objects, we describe firstly the left object. If many objects belonging to the still life are presented in an interior, they are described by enumeration, without figure-ground relations (ex. - “there are jugs, a bottle, a cup on the table”).

The last stage is text production. There is a base with morphological properties of used words. We build template constructions with number agreement control.

The text production starts by choosing a word characterizing the picture. It depends on the strategy. We can choose one of the words “landscape”, “interior” “still life”.

After that the text production bases on the patterns belonging to one of these strategies.

The example - “Interior”.

The foreground

Each of the following groups (except 0) can be omitted.

0) The Main Object A1 (it is described by the pattern K0) - 1) the upper objects, 2) the front objects, 3) the back objects, 4) the left objects 5) the right objects. The structure of the groups 1) - 5) can be various. There are the possible variants:

Case 1

If there is only one object closed to A1 (or a group of the objects building a set), each group 1) - 5) is described by the patterns:

Group 1) - K1; Group 2) - K2; Group 3) - K3; Group 4) - K4; Group 5) - K5.

Case 2

If the groups 1), 2), 3) contain two objects closed to A1 not building a set, they are described by following patterns...

Case 3

If the groups 1) - 5) contain two objects (not building a set) at the different distance from A1, they are described by following patterns...

The groups 2) - 5) can be supplemented with upper objects. In this case they are described by following patterns...

Below are some patterns:

Conventional signs:

A - object

N - name of the object describing in this sequence

I - name of the Main Object A1

A(N) - object describing in this sequence

At the first line we write down the number of the pattern and some comments. At the second line there is the template construction reflecting the order of the produced sequence parts. In square brackets there is a optional part (it is one of the following words: “great”, “small”, “another”, “one more”). In round brackets we write down the morphological properties of a word. It is the number, the choice of the number depends on the situation (we can describe a set or a single object). The next lines contain the realization of the template constructions. The slash means the variation, it depends on the

describing situation on the picture. Some patterns are compound, they contain some mutually incompatible variants (ex. pattern K1).

K0 (the Main Object A1)
PP V [Adj] Art N (pl./sg.)
PP : in the center / on the left / on the right
V: there is (if N - sg.) / there are (if N - pl.)
Art: a (if N - sg.)

K1 (A1 and A(N) touch each other)
Art1 [Adj] N (pl./sg.) V Prep [Adj] Art2 I (pl./sg.)
Art1: a (if N - sg.)
V: is (if N - sg.) / are (if N - pl.)
Prep: on
Art2: the
or
K1 (A1 - a set, A(N) touches one of the objects of I)
Art1 [Adj] N (sg./pl.) V Prep Pron Art2 [Adj] I (pl.)
Art1: a (if N - sg.)
V: is (if N - sg.) / are (if N - pl.)
Prep: on
Pron: one of
Art2: the
or
K1 (A(I) and A(N) do not touch each other)
Art1 [Adj] N (pl./sg.) V Prep [Adj] Art2 I (pl./sg.)
Art1: a (if N - sg.)
V: hangs (if N - sg.) / hang (if N - pl.)
Prep: over
Art2: the

K2 (A(N) is before A1)
V Art1 [Adj] N (pl./sg.) Prep Art2 [Adj] I (pl./sg.)
V: there is (if N - sg.) / there are (if N - pl.)
Art: a (if N - sg.)
Prep: before
Art2: the

K3 (A(N) is behind A1)
V Art1 [Adj] N (pl./sg.) Prep Art2 [Adj] I (pl./sg.)
V: there is (if N - sg.) / there are (if N - pl.)
Art: a (if N - sg.)
Prep: behind

Art2: the

K4 (A(N) is at the left of A1)
Art1 [Adj] N(pl./sg.) V Prep Art2 [Adj] I (Gen; pl./sg.)
Art1: a (if N - sg.)
Prep: at the left of
V: is (if N - sg.) / are (if N - pl.)
Art2: the

K5 (A(N) is at the right of A1)
Art1 [Adj] N(pl./sg.) V Prep Art2 [Adj] I (Gen; pl./sg.)
Art1: a (if N - sg.)
Prep: at the right of
V: is (if N - sg.) / are (if N - pl.)
Art2: the

There is an example of a generalized text. The picture - "The bedroom" of Van Gogh, 1889.

I see an interior. On the right there is a bed. There is a door before the bed. There is a rack behind the bed. Shirts are on the rack. A chair is at the left of the bed. Pictures are at the right of the bed. There are another pictures over the pictures.

On the left there is a door. There is a chair before the door. There is a towel behind the door.

On the left there is a table. Jugs, bottles, a glass, a loaf are on the table. There is a window over the table.

On the left there is a mirror.

On the right there is a picture.

The model is rather simple, we do not take into account any factors forming the salience of an object (ex. the use of an object in a picture influences its description). The model generates texts for static pictures. In the near future we will take into account more factors influencing the salience of an object.

References

Kobozeva I. How we describe the space we see: the problem of "orientation conflict", Computational Linguistics and its Applications. International Workshop, Kazan, May 31 - June 4, 1995, Proceedings.

Porzel R. Changing the Point of View and the Linearization Strategy within Spatial Descriptions: Modeling the Linearization Processes with Granular Representations, Heidelberg; Mannheim SFB 245, 1996.