Generating Referring Expressions using Perceptual Groups

Kotaro Funakoshi[1], Satoru Watanabe[1],
Naoko Kuriyama[2], and Takenobu Tokunaga[1]

[1] Department of Computer Science, Tokyo Institute of Technology
Tokyo Meguro Ôookyama 2-12-1, Japan
{koh,satoru_w,take}@cl.cs.titech.ac.jp
[2] Department of Human System Science, Tokyo Institute of Technology
Tokyo Meguro Ôookyama 2-12-1, Japan
kuriyama@hum.titech.ac.jp

Abstract. Past work of generating referring expressions mainly utilized attributes of objects and binary relations between objects to distinguish the referent from other objects. However, such an approach does not work well when there is no distinctive attribute among objects. To overcome this limitation, this paper proposes a method utilizing the perceptual groups of objects and n-ary relations among them. With the proposed method, an expression like "the leftmost ball in the left cluster of three balls" can be generated. The key is to identify groups of objects that are naturally recognized by humans. We conducted psychological experiments with 42 subjects to collect referring expressions in such situations, and built a generation algorithm based on the results. The evaluation using another 23 subjects showed that the proposed method could effectively generate proper referring expressions.

1 Introduction

Generating referring expressions is one of the important research issues of natural language generation, and many researchers have studied it [1–9].

Most past work [1–8] makes use of attributes of an intended object (the target) and binary relations between the target and others (distractors) to distinguish the target from distractors. Therefore, these methods cannot generate proper referring expressions in situations where no significant surface difference exists between the target and distractors, and no binary relation is useful to distinguish the target. Here, a *proper* referring expression means a concise and natural linguistic expression enabling hearers to distinguish the target from distractors.

For example, consider indicating object *b* to person *P* in the situation shown in Fig. 1. Note that person *P* does not share the label information such as *a* and *b* with the speaker. Because object *b* is not distinguishable from objects *a* or *c* by means of their appearance, one would try to use a binary relation between object *b* and the table, i.e., "*A ball to the right of the table*". [3] However, "*to the right of*" is not a discriminatory

[3] In this paper, we assume that all participants share the appropriate reference frame[10].

relation, for objects a and c are also located to the right of the table. Using a and c as a reference object instead of the table does not make sense, since a and c cannot be uniquely identified because of the same reason that b cannot be identified.

Van der Sluis and Krahmer [9] proposed using gestures such as pointing in situations like those shown Fig. 1. However, pointing and gazing are not always available depending on the positional relation between the speaker and the hearer.

In the situation shown in Fig. 1, a speaker can indicate object b to person P with a simple expression "the front ball" without using any gesture. In order to generate such an expression, one must be able to recognize the salient perceptual group of the objects and use the n-ary relative relations in the group. [4]

In this paper, we propose a method of generating referring expressions that utilizes n-ary relations among members of a group. Our method recognizes groups by using Thórisson's algorithm [11].

Although there are several types of relations in groups other than positional relation, such as size, e.g., "the *biggest* one", we focus on positional relations in this paper.

Speakers often refer to multiple groups in the course of referring to the target. In these cases, we can observe two types of relations: the *intra-group relation* such as "the front two *among* the five near the desk", and the *inter-group relation* such as "the two *to the right of* the five". We define that a subsumption relation between two groups is an intra-group relation.

In what follows, Sect. 2 explains the experiments conducted to collect expressions in which perceptual groups are used. The proposed method is described in Sect. 3, and the evaluation is described in Sect. 4. Finally, we conclude the paper in Sect. 5.

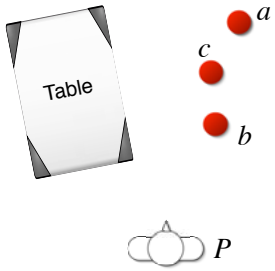

Fig. 1. An example of problematic situations

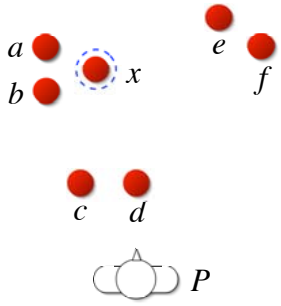

Fig. 2. A visual stimulus of the experiment

[4] Although Krahmer *et al*. claim that their method can handle n-ary relations [8], they provide no specification. We think their method cannot directly handle situations we discuss here.

2 Data Collection

We conducted a psychological experiment with 42 Japanese undergraduate students to collect referring expressions in which perceptual groups are used. In order to evaluate the collected expressions, we conducted another experiment with a different group of 44 Japanese undergraduate students. There is no overlap between the subjects of those two experiments. Details of this experiment are described in the following subsections.

2.1 Collecting Referring Expressions

Method Subjects were presented 2-dimensional bird's-eye images in which several objects of the same color and the same size were arranged and the subjects were requested to convey a target object to the third person drawn in the same image. We used 12 images of arrangements. An example of images presented to subjects is shown in Fig. 2. Labels $a, \ldots, f, x$ in the image are assigned for purposes of illustration and are not assigned in the actual images presented to the subjects. Each subject was asked to describe a command so that the person in the image picks a target object that is enclosed with dotted lines. When a subject could not think of a proper expression, she/he was allowed to abandon that arrangement and proceed to the next one. Referring expressions designating the target object were collected from these subjects' commands.

Analysis We presented 12 arrangements to 42 subjects and obtained 476 referring expressions. Twenty eight cases were abandoned in the experiment. Observing the collected expressions, we found that starting from a group with all of the objects, subjects generally narrow down the group to a singleton group that has the target object. Therefore, a referring expression can be formalized as a sequence of groups (SOG) reflecting the subject's narrowing down process.

The following example shows an observed expression describing the target x in Fig. 2 with the corresponding SOG representation below it.

"*hidari oku ni aru mittu no tama no uti no iti-ban migi no tama*."
(the rightmost ball among the three balls in the back-left side)

SOG: $[\{a, b, c, d, e, f, x\}, \{a, b, x\}, \{x\}]$ [5]
where
$\{a, b, c, d, e, f, x\}$ denotes all objects in the image (total set),
$\{a, b, x\}$ denotes the three objects in the back-left side, and
$\{x\}$ denotes the target.

Since narrowing down starts from the total set, the SOG representation starts with a set of all objects and ends with a singleton group with the target. Translating the collected referring expressions into the SOG representation enables us to abstract and classify the expressions. On average, we obtained about 40 expressions for each arrangement, and classified them into 8.4 different SOG representations. The summary of collected data is shown in Table 1.

Although there are two types of relations between groups as we mentioned in Sect. 1, the expressions using only intra-group relations made up about 80% of the total.

[5] We denote a SOG representation by enclosing groups with square brackets.

Table 1. Summary of the collected data

Arrangement ID	1	2	3	4	5	6	7	8	9	10	11	12	Average
Number of expressions obtained	41	40	41	41	42	37	42	32	42	41	41	36	39.7
Number of different SOGs	5	6	8	8	6	12	4	15	4	11	5	17	8.4

2.2 Evaluating the Collected Expressions

Method Subjects were presented expressions collected in the experiment described in Sect. 2.1 together with the corresponding images, and were requested to indicate objects referred to by the expressions. The presented images are the same as those used in the previous experiment except that there are no marks on the targets. At the same time, subjects were requested to express their confidence in selecting the target, and evaluate the conciseness, and the naturalness of the given expressions on a scale of 1 to 8.

Because the number of expressions that we could evaluate with subjects was limited, we chose a maximum of 10 frequent expressions for each arrangement. The expressions were chosen so that as many different SOG representations were included as possible. If an arrangement had SOGs less than 10, several expressions that had the same SOG but different surface realizations were chosen. The resultant 117 expressions were evaluated by 44 subjects. Each subject evaluated about 32 expressions.

Analysis Discarding incomplete answers, we obtained 1,429 evaluations in total. 12.2 evaluations were obtained for each expression on average.

We measured the quality of each expression in terms of an *evaluation value* that is defined in (1). This measure is used to analyze what kind of expressions are preferred and to set up a scoring function (5) for machine-generated expressions as described in Sect. 3.

$$(evaluation\ value) = (confidence) \times \frac{(naturalness) + (conciseness)}{2} \qquad (1)$$

According to our analysis, the expressions with only intra-group relations obtained high evaluation values, while the expressions with inter-group relations obtained lower evaluation values. We provide a couple of example expressions indicating object x in Fig. 2 to contrast those two types of expressions below.

- without inter-group relations (i.e., with intra-group relations only)
  - "the rightmost ball among the three balls in the back-left side"
- with inter-group relations
  - "the ball behind the two front balls"

In addition, expressions explicitly mentioning all the objects obtained lower evaluation values. Considering these observations, we built a generation algorithm using only intra-group relations and did not mention all the objects explicitly.

The summary of the analysis is shown in Table 2. The first column "w/o inter-group" shows the data concerning the expressions with intra-group relations only. The second column "w/ inter-group" shows the data concerning the expressions with inter-group relations.

Table 2. Statistics of the human evaluation

		w/o inter-group	w/ inter-group	total
Number of expressions		86	31	117
Accuracy	: Range (%)	9 - 100	0 - 100	0 - 100
	: Average (%)	93.51	70.02	87.29
	: Std. Dev.	16.28	35.04	23.61
Evaluation Value	: Range	11.66 - 55.54	10.49 - 49.54	10.49 - 55.54
	: Average	34.40	25.16	31.95
	: Std. Dev.	10.14	11.42	10.89
Confidence	: Range	3.93 - 7.75	3.36 - 7.36	3.36 - 7.75
	: Average	6.36	5.59	6.15
	: Std. Dev.	0.91	1.17	1.01
Briefness	: Range	2.85 - 7.36	2.25 - 7.00	2.25 - 7.36
	: Average	5.53	4.59	5.28
	: Std. Dev.	1.00	1.26	1.09
Naturalness	: Range	2.75 - 7.18	2.33 - 6.18	2.33 - 7.18
	: Average	5.09	4.09	4.83
	: Std. Dev.	1.07	1.28	1.17

Among these expressions, we selected those with which the subjects successfully identified the target with more than 70% accuracy. The selected expressions are used to build a generation algorithm. One might think this threshold is too low. However, since the average number of evaluations for each expression is not so large, deleting incomplete evaluations further reduces the number of evaluations, and thus decreases the reliability. 70% is a compromised threshold value. These expressions are used to extract parameters of our generation algorithm in the next section.

3 Generating Referring Expressions

Given an arrangement of objects and a target, our algorithm generates referring expressions by the following four steps:

Step 1: enumerate perceptual groups based on the proximity between objects
Step 2: generate the SOG representations by combining the groups
Step 3: calculate the scores of each SOG representation
Step 4: translate the SOG representations into linguistic expressions

In the rest of this section, we illustrate how these four steps generate referring expressions in the situation shown in Fig. 2.

Step 1: Generating Perceptual Groups. To generate perceptual groups from an arrangement, Thórisson's algorithm [11] is adopted.

Given a list of objects in an arrangement, the algorithm generates groups based on the proximity of the objects and returns a list of groups. Only groups containing the target, that is x, are chosen because we handle intra-group relations only as mentioned

before, and that implies that all groups mentioned in an expression must include the target. Then, the groups are sorted in descending order of the group size. Finally a singleton group consisting of the target is added to the end of the list if such a group is missing in the list. The resultant group list, GL, is the output of Step 1.

For example, the algorithm recognizes the following groups given the arrangement shown in Fig. 2:

$$\{\{a, b, c, d, e, f, x\}, \{a, b, c, d, x\}, \{a, b, x\}, \{c, d\}, \{e, f\}\}.$$

After filtering out the groups without the target and adding a singleton group with the target, we obtain the following list:

$$GL = \{\{a, b, c, d, e, f, x\}, \{a, b, c, d, x\}, \{a, b, x\}, \{x\}\}. \tag{2}$$

Step 2: Generating the SOG Representations. In this step, the SOG representations introduced in Sect. 2 are generated from the GL of Step 1, which generally has a form like (3), where G_i denotes a group, and G_0 is a group of all the objects. Here, we narrow down the objects starting from the total set (G_0) to the target ($\{x\}$).

$$GL = \{G_0, G_1, \ldots, G_{m-2}, \{x\}\} \tag{3}$$

First, given a group list GL, all possible SOGs are generated. From a group list of size m, 2^{m-2} SOG representations can be generated since G_0 and $\{x\}$ should be included in the SOG representation. For example, from a group list of $\{G_0, G_1, G_2, \{x\}\}$, we obtain four SOGs: $[G_0, \{x\}]$, $[G_0, G_1, \{x\}]$, $[G_0, G_2, \{x\}]$, and $[G_0, G_1, G_2, \{x\}]$.

Second, among these generated SOG representations, those that contain more than four groups are discarded at this stage for the sake of conciseness. This filtering was introduced considering the observation of the collected data.

For example, one of the SOG representations generated from list (2) is

$$[\{a, b, c, d, e, f, x\}, \{a, b, x\}, \{x\}]. \tag{4}$$

Step 3: Calculating Scores. This step calculates the score of SOG representations. Currently, the score is calculated only based on the SOG representation, that is, features of linguistic expressions such as phrase length, are not considered.

The total score of an SOG representation is calculated by averaging the scores given by functions f_1 and f_2 whose parameters are dimension ratios between two consecutive groups as given in (5), where n is the number of groups in the SOG representation.

$$score(SOG) = \frac{1}{n-1}\left\{\sum_{i=0}^{n-3} f_1\left(\frac{dim(G_{i+1})}{dim(G_i)}\right) + f_2\left(\frac{dim(\{x\})}{dim(G_{n-2})}\right)\right\} \tag{5}$$

The dimension of a group dim is defined as the average distance between the centroid of the group and that of each object. The dimension of the singleton group $\{x\}$ is defined as a constant value. Because of this idiosyncrasy of the singleton group $\{x\}$ compared to other groups, function f_2 was introduced separately from function f_1 even though both functions represent the same concept as described below.

We assume that, when a speaker tries to narrow down an object group from G_i to G_{i+1}, there is an optimal ratio between the dimensions of G_i and G_{i+1}. In other words, narrowing down a group from a very big one to a very small one might cause hearers to become confused. For example, consider the following two expressions that both indicate object x in Fig. 2. Hearers would prefer (A) to (B) though (B) is simpler than (A).

(A) "the rightmost ball among the three balls in the back-left side"
(B) "the fourth ball from the left"

The optimal ratio between two groups, and that from a group to the target were found through the quadratic regression analysis of data collected in the experiment described in Sect. 2.2. Functions f_1 and f_2 are the two regression curves found through the analysis representing correlations between dimension ratios and values calculated based on human evaluation as in (1).

Step 4: Generating Linguistic Expressions. In the last step, the SOG representations are translated into linguistic expressions. Since Japanese is a head-final language, the order of linguistic expressions for groups are retained in the final linguistic expression for the SOG representation. That is, an SOG representation $\{G_0, G_1, \ldots, G_{n-2}, \{x\}\}$ can be achieved as shown in (6), where $E(X)$ denotes a linguistic expression for X, $R(X, Y)$ denotes a relation between X and Y, and '+' is a string concatenation operator.

$$E(G_0) + E(R(G_0, G_1)) + E(G_1) + \ldots + E(R(G_{n-2}, \{x\})) + E(\{x\}) \quad (6)$$

As described in Sect. 2.2, expressions that explicitly mention all the objects obtain lower evaluation values, and expressions using intra-group relations obtain high evaluation values. Considering these observations, our algorithm does not use the linguistic expression corresponding to all the objects, that is $E(G_0)$, and only uses intra-group relations for $R(X, Y)$.

Possible expressions of X are collected from the experimental data in Sect. 2.1, and the first applicable expression is selected when realizing a linguistic expression for X, i.e., $E(X)$.

For example, the SOG representation (4) is realized as follows.

$$[\{a, b, c, d, e, f, x\}, \{a, b, x\}, \{x\}]$$

$$\rightarrow E(R(\{a, b, c, d, e, f, x\}, \{a, b, x\})) + E(\{a, b, x\}) + E(R(\{a, b, x\}, \{x\})) + E(\{x\})$$

→ *"hidari oku no"*+*"mittu no tama"*+*"no uti no migihasi no"*+*"tama"*
(in the back-left) (three balls) (rightmost ... among) (ball)

Note that there is no mention of all the objects, $\{a, b, c, d, e, f, x\}$, in the linguistic expression.

Table 3. Summary of evaluation

	Accuracy (%)	Naturalness	Conciseness	Confidence	Eval. val.	Agreement (%)
Human-1	87.3	4.82	5.27	6.14	32.0	N/A
Human-2	97.0	5.05	5.49	6.38	34.2	N/A
System-A	91.0	5.60	6.25	6.32	40.1	53.3
System-B	88.4	5.09	5.65	6.25	35.2	45.0
System-Average	89.2	5.24	5.82	6.27	36.6	46.7

4 Evaluation

We implemented the algorithm described in Sect. 3, and evaluated the output with 23 undergraduate students. The subjects were different from those of the previous experiments but were of the same age group, and the experimental environment was the same. The evaluation of the output was performed in the same manner as that of Sect. 2.2.

The results are shown in Table 3. "Human-1" shows the average values of all expressions collected from humans as described in Sect. 2.2. "Human-2" shows the average values of expressions by humans that gained more than 70% in accuracy in the same evaluation experiment. Our algorithm tries to emulate the expression of "Human-2", thus this would be the baseline of the algorithm.

"System-A" shows the average values of expressions generated by the algorithm for the 12 arrangements used in the data collection experiment described in Sect. 2.1. The algorithm generated 18 expressions for the 12 arrangements, which were presented to each subject in random order for evaluation.

"System-B" shows the average values of expressions generated by the algorithm for 20 randomly generated arrangements that generate at least two linguistic expressions. The algorithm generated 48 expressions for these 20 arrangements, which were evaluated in the same manner as that of "System-A".

"System-Average" shows the micro average of expressions of both "System-A" and "System-B".

"Accuracy" shows the rates at which the subjects could identify the correct target objects from the given expressions. Comparing the accuracies of "Human-2" and "System-A", we find that the algorithm generates very good expressions. Moreover, the algorithm is superior to human in terms of "Naturalness" and "Conciseness". However, this result should be interpreted carefully. Further investigation of the expressions revealed that humans often sacrificed naturalness and conciseness in order to describe the target as precisely as possible for complex arrangements.

The last column, "Agreement", shows to what extent the scores of expressions given by the algorithm conform with the human evaluation. The agreement is calculated as follows. First, the generated expressions are ordered according to the algorithm's score given by (5) in Sect. 3 and the human evaluation given by (1) in Sect. 2.2. All binary order relations between two expressions are extracted from these two ordered lists of expressions respectively. The agreement is defined as the ratio of the same binary order relations among the number of all binary orders. We find that the current scoring function does not conform with the human evaluation very well.

5 Concluding Remarks and Future Work

This paper proposed an algorithm that generates referring expressions using perceptual groups and n-ary relations among them. The algorithm was built on the basis of the analysis of expressions that were collected through linguistic experiments. The performance of the algorithm was evaluated by 23 subjects and it generated promising results.

In the following, we look at future work to be done.

Recognizing Lines: Thórisson's algorithm [11] cannot recognize objects in linear arrangement as a group, although such an object arrangement is quite salient for humans. This is one of the reasons for the disconformity of the evaluations between those of the algorithm and those of the human subjects.

For example, in the arrangement shown in Fig. 3, Thórisson's algorithm will recognize groups $G1$, $G2$, and $G4$ but not group $G3$ because the distance between objects x and c is a little bit longer than other distances between objects. However, a line formed by group $G3$ is salient for humans, and it would be preferred to use $G3$ to generate expressions such as "the second one from the left among the four balls in back".

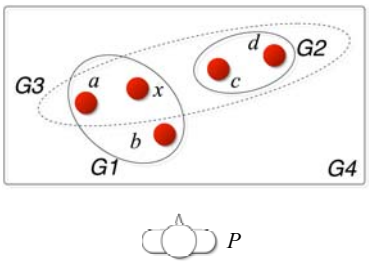

Fig. 3. An example arrangement forming a line

Using Relations Other Than Positional Relations: In this paper, we focused on positional relations of perceptual groups. Other relations such as degrees of colors and sizes should be treated in the same manner.

Designing a Better Scoring Method: As shown by the evaluation in Sect. 4, our scoring method described in Sect. 3 does not conform with the human evaluation. The method uses only the dimension ratio of groups in the course of the narrowing down process. This would be an effective factor for generating appropriate referring expressions but not necessarily the primary one. Further research is required to explore other factors to be incorporated into the scoring method.

Integrating with the Conventional Methods: In this paper, we focused on the limited situation where inherent attributes of objects are useless, but this is not the case in

general. The algorithm integrating the conventional attribute-based methods and the proposed method should be investigated to achieve the end goal.

A possible direction would be enhancing the algorithm proposed by Krahmer *et al*. [8]. They formalize an object arrangement (*scene*) as a labeled directed graph in which vertices model objects and edges model attributes and binary relations, and regard content selection as a subgraph construction problem. Their algorithm performs searches directed by a cost function on a graph to find a subgraph that has no isomorphic subgraphs on the same graph.

By considering a perceptual group as an ordinary object, their algorithm is applicable. However, introducing perceptual groups as vertices makes it difficult to design the cost function. A well-designed cost function is indispensable for generating concise and comprehensible expressions. Otherwise, an expression like "a ball in front of a ball in front of a ball" for the situation shown in Fig. 1 would be generated.

References

1. Appelt, D.E.: Planning English referring expressions. Artificial Intelligence **26** (1985) 1–33
2. Dale, R., Haddock, N.: Generating referring expressions involving relations. In: Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics(EACL'91). (1991) 161–166
3. Dale, R.: Generating referring expressions: Constructing descriptions in a domain of objects and processes (1992) MIT Press, Cambridge.
4. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science **19** (1995) 233–263
5. Heeman, P., Hirst, G.: Collaborating referring expressions. Computational Linguistics **21** (1995) 351–382
6. Krahmer, E., Theune, M.: Efficient context-sensitive generation of descriptions (2002) In Kees van Deemter and Rodger Kibble, editors, Information Sharing: Givenness and Newness in Language Processing. CSLI Publications, Stanford, California.
7. van Deemter, K.: Generating referring expressions: Boolean extensions of the incremental algorithm. Computational Linguistics **28** (2002) 37–52
8. Krahmer, E., van Erk, S., Verleg, A.: Graph-based generation of referring expressions. Computational Linguistics **29** (2003) 53–72
9. van der Sluis, I., Krahmer, E.: Generating referring expressions in a multimodal context: An empirically oriented approach (2000) Presented at the CLIN meeting 2000, Tilburg.
10. Levinson, S.C., ed.: Space in Language and Cognition. Cambridge University Press (2003)
11. Thórisson, K.R.: Simulated perceptual grouping: An application to human-computer interaction. In: Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society. (1994) 876–881