社団法人 情報処理学会 研究会報告
IPSJ•SIG•Technical•Reports

2007-NL-180 (11) 2007/7/24

参照表現における知覚的群化について

スパンガーフィリップ† 徳永健伸†

† 東京工業大学大学院情報理工学研究科
E-mail: †{philipp,take}@cl.cs.titech.ac.jp

あらまし 参照表現の理解は人間とコンピュータの対話にとって重要な研究テーマである．本稿では，従来提案されていた参照物体の群を利用した参照表現の理解モデルを拡張し，より多様な参照表現を扱えるように，モデルを一般化する．従来モデルでは,「左にある三つの玉のうちの右の玉」のように対象を特定する線形な絞り込み過程に注目し，これをモデル化している．たしかに，実験の結果，このような表現を人間が多用する傾向があることが判っている．しかし，人間が生成する参照表現には，線形の絞り込み過程では説明できない例もあり，線形の絞り込みでは自然な参照表現を生成できない場合もある．本稿では,「机の前と椅子の後ろにある玉」のような非線形性 (包含関係以外の対象物体の群の関係) を含む参照表現も扱えるモデルを提案する．英語，ドイツ語，フランス語について実験をおこない，その有効性を確認した．また，このモデルの言語の依存性についても議論する．

キーワード 参照表現，知覚的群化，言語依存性

On perceptual grouping of reference expressions

Philipp SPANGER† and TOKUNAGA Takenobu†

† Department of Computer Science, Tokyo Institute of Technology
E-mail: †{philipp,take}@cl.cs.titech.ac.jp

Abstract Understanding of reference expressions is critical in human-agent interaction through natural language. This paper extends and generalizes an existing formal model of reference expressions involving perceptual grouping, to take account of the wide variety of human-produced reference expressions. The previous model can only represent a linear process of narrowing down of the referent; e.g.: "the right ball of the three balls in the left". While psychological experiments indicate a general tendency towards humans using expressions of this type, it is far from the only or even always the most natural one. In particular, our model can account for non-linear expressions (relations of sets of objects, other than exclusively the subsumption-relation) that are used by humans, like: "the ball in front of the desk and behind the chair". We conducted an experiment to collect data of reference expressions in English in order to evaluate the proposed model. Our proposed model yielded an increase in both coverage and accuracy of referent identification. An algorithm is then outlined for the application of this model to other languages.

Key words reference expressions, perceptual grouping, language-independent systems

1. Introduction

Recently, the development of a multitude of research areas like speech recognition, robotics, etc. has enabled important progress in developing agents aimed at real-world interaction with humans. A central objective in improving agents' capabilities of interaction with humans is to improve their natural language understanding. A fundamental type of human expression – in particular in task-oriented dialogue – are reference expressions. This type of expression is a linguistic entity used to discriminate a specific object from its environment and the rest of the world. Thus, agents' capabilities to handle this type of linguistic expressions correctly is an important part of increasing human-agent interaction capabilities.

Reference expressions are to a large degree multi-modal; i.e. they include exophoric expressions such as "this one" or "that" in connection with some gestures (e.g.; pointing). It is clear, a fuller model of reference expressions must be a multi-modal model including an account of these different

channels and how they combine (e.g.; Kranstedt et al, 2006 [1]). As a first preliminary step towards this aim, we intend to generalize a current model of reference expressions limited to the linguistic channel, as a basis for future application in a multi-modal environment. Hence, in this paper reference expression refers to a single-channel linguistic expression which moreover includes no anaphora and is functioning as a full description for identifying objects in the world such as "the blue ball in front of the table".

Previous research has underlined the importance of perceptual grouping in understanding and generating reference expressions [2, 3, 5]. Perceptual grouping is defined as the human ability to recognize similar objects, or objects in close proximity to each other. Effective understanding of human reference expressions requires recognition of similar or proximal objects, i.e., perceptual grouping, and requires making use of n-ary relations among objects in each recognized group. Research based on this understanding has produced comparably good results in both understanding [2] and generation of reference expressions [3].

While this general approach has proven valuable in both the understanding and generation of this type of expression, it has been hampered – both theoretically and in practice – by a strong limitation on the type and structure of expression. That is, it has been assumed reference expressions exclusively apply a *linear process of narrowing down of the referent* (represented by a 'Sequence of groups- representation' (SOG) in [2] and in its generalised form in [3]). However, this means other relations between sets of perceptual groups (appearing in reference expressions) like intersection or subtraction cannot be represented.

Data of experiments in several languages (Japanese, English, German, French) indicate that while the overwhelming majority of expressions (in all four languages) is based on this type of process, it is far from the only or even always the most natural one for humans. In particular, humans are capable and in some cases prefer to refer to different types of relations between sets of similar objects using either intersection or subtraction. In certain cases this simplifies the expression significantly or is more natural.

For example, in Figure 1 the subject forms an intersection of the group of all balls "behind the round table" and the group of balls "in front of the square table". The set that is the result of this intersection is the one ball the subject intended to refer to. This is one example of a process of referring, that cannot be represented in the previous model.

Hence, in order to develop the promising framework of perceptual grouping in reference expressions, it is necessary to generalize the existing model such that it can accomodate these more complex cases. This paper tackles this task. This

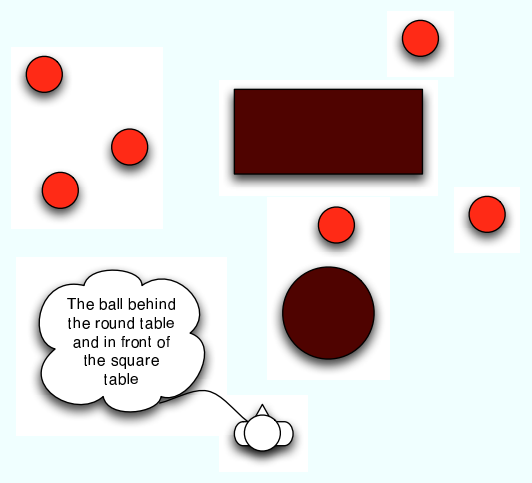

図 1　積集合を用いた例
Fig. 1　An example of an expression using intersection

will make a contribution to increasing our understanding of reference expressions and provide a general theoretical model of this type of expression.

As in [2], we consider here that understanding reference expressions consists of two stages: (a) semantic analysis, i.e., analyzing expressions to extract semantic information, and (b) referent identification, i.e., discriminating referents by using extracted information. Below, we describe the proposed generalized model of reference expressions and how it handles the more complex cases the previous model could not deal with. We then explain some modifications of the algorithm for perceptual grouping proposed in [2], as necessitated by the more general model proposed in this paper. We will discuss the collected data (in English as well as other languages) and the implementation of the proposed model in a simple prototype that yielded an increase in both coverage and referent identification. Finally, we will give an overview of future work on this topic.

2. A formal model of reference expressions

The majority of previous work on reference to a target-object among other distractors, (e.g.; Appelt, 1985; Dale and Haddock, 1991; Dale, 1992; Dale and Reiter, 1995; Heeman and Hirst, 1995; Horacek, 1997; Krahmer and Theune, 2002; van Deemter, 2002; Krahmer et al., 2003) utilized attributes of the target and binary relations between the target and distractors, using surface differences of the objects. However, in case no significant surface difference and no binary relation useful to distinguish the target from the distractors exists, such methods could not generate a natural linguistic expression enabling hearers to identify the target.

To solve this insufficiency, Funakoshi et al. (2004) proposed a method of generating Japanese reference expressions

that utilizes n-ary relations among members of a group. In this framework, they presented an intermediate "Sequence of groups"(SOG) - representation, in order to capture the linear process of narrowing down of the referent.

However, their framework only dealt with the limited situations where only homogeneous objects are randomly arranged (as in Figure 1). Thus, their method could handle only spatial n-ary relation, and could not handle attributes and binary relations between objects which have been the main concern of the past research.

2.1 The (extended) SOG representation and its limitations

(Funakoshi et al., 2004) assumed a situation with randomnly-arranged homogenous objects and focused exclusively on spatial subsumption relations between consecutive groups. Thus, the intermediate representation they proposed between a reference expression and the situation that is referred to by the expression, did not explicitly denote relations between groups in the original SOGs. (as shown below).

$SOG : [G_0, G_1, ...G_n]$
$G_i : a\ group$

In order to take into account other types of relations between groups, (Funakoshi et al., 2006) proposed then an extended SOG representation where types of relations are explicitly denoted as shown below.

$SOG : [G_0 R_0 G_1 R_1 ... G_n]$
$G_i : a\ group$
$R_i : relation\ between\ G_i\ and\ G_{i+1}$

In this extended model, two types of relations between groups were accounted for: intra-group relations and inter-group relations. Of course, for any intra-group relation, by definition, G_i subsumes G_{i+1}, that is, $G_i \supset G_{i+1}$. Intra-group relations are further classified into subcategories according to the feature used to narrow down G_i to G_{i+1}.

In this model, in case R_i is an inter-group relation, G_i and G_{i+1} are mutually exclusive, that is, $G_i \cap G_{i+1} = \emptyset$. However, this leaves out cases of other inter-group relations, in particular other combinations of perceptual groups like intersection ($G_i \cap G_{i+1}$) or subtraction ($G_i \setminus G_{i+1}$). The necessity of incorporating this type of expression can be demonstrated by, for example, Figure 2 (example from the collected expressions).

2.2 The COG (combination of groups) representation

The COG representation is a generalisation and extension of the (extended) SOG-representation. Its flexible order of grouping ("linear" or "non-linear") better captures the natural variety of human reference expressions. It includes the SOG-representation as a special case.

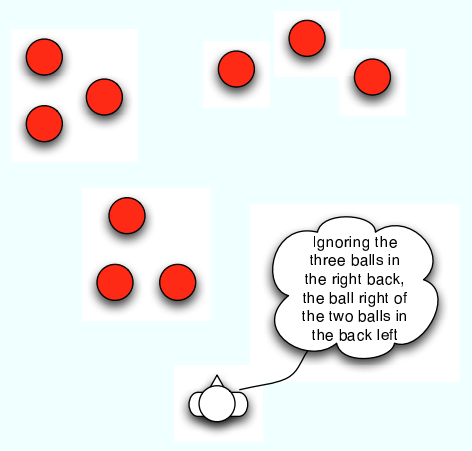

図 2　差集合を用いた例
Fig. 2　An example of an expression using subtraction

The initial SOG-model has been extended to the case of different relations among sets, which arise in more complex environments with a number of different objects. However, both the extension as well as the initial model share the same weakness, in not accounting for reference expressions that include in some form a non-linear process of narrowing down the referent. Thus, in order to improve the model, we implemented the generalisation based on the simpler earlier model in order to demonstrate the validity of the proposed generalisation. We note, this generalisation is applicable as well for the extended SOG-model.

Watanabe et. al (2006) point out that in their method, "most errors in semantic analysis are due to non-linearity of referring". This is because the SOG-model presupposes linearity in referring and thus cannot handle these cases.

We conducted a data collection experiment in 3 different languages; English, French and German. We presented subjects with images prepared by Watanabe et. al (2006), then collected the expressions the subjects deemed appropriate to distinguish a specified ball. 12 subjects took part in the collection of data in English, slightly more than in French (8 subjects) and German (10 subjects). As far as possible, we ensured only native speakers took part in the experiment. Analysis of the collected data indicated two general cases of "non-linearity" in referring; use of either the intersection or the subtraction - relation.

As the name "*Sequence* of groups" indicates, this model has a "flat" structure; i.e. only accounts for simple relations to the immediately preceding group. In contrast, the proposed COG-model allows an internal structure; i.e. a reference to any previous group or combination of groups. Thus this model can correctly represent the cases where the SOG fails (i.e.: subsumption and intersection relation). The

general model can be represented as follows:

$COG : [G_0 R_0 G_1 .. [G_i R_i .... G_j] G_j .... G_n]$
G_i : *a group*
R_i : *relation between the preceding and succeeding combination of groups*

Thus, theoretically an arbitrary level of complexity of grouping-(reference) expressions can be represented in this model.

We note the extended SOG-model by (Funakoshi et al., 2006) is a special case of the proposed COG-model, in the case that all relations are restricted only to sumsumption-relations and thus only relations to immediately preceding groups are allowed. It is a subset of the proposed COG-model and thus incorporated into the model. In the following, based on our proposed model we provide the analysis of an example, that cannot be accounted for by the previous model (displayed in Figure 3).

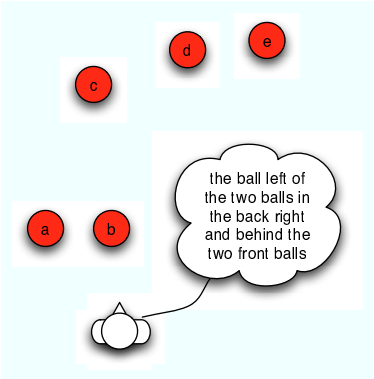

図 3　解　析　例
Fig. 3　Example analysis

2. 3 Example analysis in COG

We present the analysis in the COG-model of the example presented in Figure 3, that as we pointed out previously the SOG-model cannot handle. We recapitulate the phrase: "the ball left of the two balls on the right and behind the two front balls". The analysis would look as follows :

$COG : [\{a,b,c,d,e\} R_0 (\{a,b,c\} R_1 \{c,d,e\}) \{c\}]$
R_0: *subsumption relation*
R_1: *intersection relation*

Thus the resulting set $\{c\}$ is a result of the intersection of the two sets $\{a,b,c\}$ and $\{c,d,e,f\}$ and the relation R_0 is a relation over the whole intersection relation.

2. 4 Implementation

Following [2], we consider the general process of reference expression understanding to consist of the following two stages (a) semantic analysis, i.e., analyzing expressions in order to extract semantic information, and (b) referent identification, i.e., uniquely recognizing referents by using the extracted semantic information.

Generally, the methods of [2] are employed for both stages, in particular in the process of perceptual grouping. However, in both stages some modification of the methods were implemented.

(a) Semantic Analysis

[2] used a simple pattern matching-technique for extracting the necessary information from the linguistic expressions, instead of full parsing. In this paper, we used the Stanford Parser (see [4]) to get a basic syntactic structure based on PCFG. We then analysed the basic syntactic structures of the user inputs and recognized that to a large degree the syntactic structure gives a good clue as to how to separate a clause into groups for extracting the required information.

We recognize this is still a partial progress to a pattern matching technique. In future a more complete analysis and comparison of different grammatical approaches should be carried out.

(b) Referent identification

A simple but conceptually important modification to [2] was implemented. The algorithm proposed in [2] only recognizes groups that are explicitly referred to, with their cardinality specified (e.g.: "the three balls in the front right"). As pointed out above, in addition to this type of perceptual grouping, humans carry out perceptual grouping "by exclusion"; i.e. "the balls right to the table and ..." . Here the subject forms a specific perceptual group G_1, without specifying a cardinality of the group.

This modification was important in particular in order to implement the more complex combinations of groups like repeated intersections and complements. In fact, the limitation to allow only perceptual groups with explicit cardinality is closely connected to the "linear" structure of the SOG-model; since neither intersection nor subtraction- relations on sets were permitted, the only way to select certain elements from a super-set is to specify its cardinality. Hence, in our generalisation to allow a wider range of set-relations, we needed to implement a concomittant generalisation of the perceptual grouping algorithm proposed in [2].

3. An outline of a language-independent algorithm

The proposal of the COG-representation is based on observations of data-collection in English, French and German. It captures the general structure of reference expressions in these languages. Of course, in order to implement a comparable system of understanding reference expressions, the possibly very significant differences in syntax have to be accounted for. From the testdata, our observation is that there is no significant fundamental difference in perceptual group-

ing that would force a fundamental revision of the algorithm proposed in this paper.

However, we found some tendencies of preference of certain types of expressions, which differed over the different languages. Generally, we observed some interesting characteristics in the collected data in the different languages. German expressions showed a significant variation in the syntactic structure of the expressions (as well as the used vocabulary), while the expressions supplied by the French subjects showed a very high degree of similarity of syntactic structure. Further study in other languages should illuminate this phenomenon and the connection between cognitive and linguistics processes.

In order to implement and test the proposed model in other languages, in particular the following modules should be prepared:

- Syntactic parser

The output of the parser should be analysed for indicators in the syntactic structure of distinct grouping.

- Information extraction-module

Based on the previous step, words/ syntactic structures indicating a particular set-relation of perceptual groups (e.g.: subtraction - in English "ignoring ..." etc.) should be identified and applied to extract the relevant information.

In our implementation of the methods of referent identification developed in [2] for Japanese, we noted there was no significant modification necessary for the application of these methods to English, other than those explained above. Thus, our data-collection experiment indicates a very universal process of perceptual grouping. English and Japanese– being two languages with significant differences in syntax– provide at least a good basis for making this hypopthesis. Our testdata in French and German confirm this hypothesis. Of course the amount of testdata is very small and thus these hypotheses need to be tested using a more comprehensive set of data.

4. Evaluation

We implemented the proposed model in Java and applied them to the expressions collected in our data-collection experiment. We then evaluated the referent identification accuracy of the proposed model.

4.1 Experiment

We carried out a data-collection experiment for English, where we provided the 12 different arrangements of balls in a 2-D bird's-eye image to the subjects (taken from the appendix in [2]). 12 subjects whose native language is English participated in the experiment over the internet. They were provided an arrangement with the choice to either input an expression they felt appropriate or to abandon this specific arrangement in case subjects were not able to think of an appropriate reference expression.

This should have produced 144 expressions, however 7 judgements were abandoned and 15 expressions were either nonsensical or obviously insufficient to identify the referent. Hence, we obtained 122 English reference expressions.

表 1　従来モデルの結果 (英語)
Table 1　Results of previous model in English

Expression	*Pop.*	*Ident.*
Total	122	77.0%
Applicable	107	83.1%

表 2　従来モデルの結果 (日本語)
Table 2　Results of previous model in Japanese

Expression	*Pop.*	*Ident.*
Total	476	78.8%
Applicable	425	84.7%

[2] referred to about 8% of Japanese collected expressions that included non-linearity of referring. In the English data collected in our experiment, we noted a 7% frequency of reference expressions that include non-linearity. This points toward very similar frequency of this type of expression. However, the amount of the English data in particular is small; hence in order to confirm this hypothesis a larger data set is necessary.

4.2 Results

We did not have the previous system (in Japanese) at our disposal. Thus we implemented the algorithm as outlined in [2] in English. This was in order to provide a baseline for our proposed enhanced model.

The result of this system is represented in Table 1 in comparison with the results of the Japanese system, displayed in Table 2. Our system in English based on the algorithm in [2] gives largely comparable results to the system of [2] for Japanese. The slight decrease in accuracy (about 2%) can be attributed in part to a lack of fine-tuning of the algorithms for perceptual grouping.

The result of the implemented system based on [2] in English is represented in Table 1. It shows that the simple implementation in English yielded a comparable result to the Japanese system, while having slightly less accuracy. This might be attributed to slight differences in implementation; e.g. setting of some parameters in the formulas of the perceptual grouping methods.

We then implemented the GOG-model and the result is represented in Table 2. This implementation of the GOG-model yielded an increase of 5.6% in comparison to the SOG-model in English. The final accuracy achieved for English was 82.6%.

表 3　提案手法の結果 (英語)

Table 3　Results of improved model in English

Expression	*Pop.*	*Ident.*
Total	122	82.6%
Applicable	114	89.2%

4. 3 Error analysis

(a) Errors in semantic analysis

There were two main types of errors in semantic analysis. One type of expression referred to a particular part of the body of the person in the picture as referent, e.g.: "the one in front of my left shoulder". The other frequent type of expression that cannot be handled by our system are expressions that refer to an action, like : "Take away the two left ones and you 'll have it now as the most left ones". There were 3 expressions of this type. This type of expression appeared in all three languages of our experiment, thus indicating it is not an isolated phenomenon. A future system should be able to handle expressions of this type involving actions .

(b) Referent identification

In the main, errors were due to reference to geometric forms–in particular lines–and our current system cannot handle any perceptual grouping involving this type of figure.

We acknowledge that this is a preliminary evaluation, as the test-data is less than a quarter of the amount of expressions collected in the other system.

5. Conclusion

We proposed a generalised model of reference expressions that seeks to capture the varied forms of reference expressions employed by humans. Our model increased both coverage and identification accuracy in comparison to the previous model in English by 5.6%. We gave some outline of how to implement this general framework in other languages.

6. Future Work

Our model has so far only been implemented in the area of understanding of reference expressions, but it should be noted that it could be extended to generation of reference expression.

The proposed model is simply a linguistic model of reference expressions in a 2-D environment. In order to increase the efficiency of human-agent communication it is necessary to incorporate other channels of communication ("multi-modality") and combine the information of these. We plan to extend and adapt the proposed model in this paper to a multi-modal environment.

文　献

[1] A. Kranstedt, A. Lking, T. Pfeiffer, H. Rieser and I. Wachsmuth (2006). Deictic object reference in task-oriented dialogue. In G. Rickheit and I. Wachsmuth (eds.): Situated Communication. pp. 155-207. Mouton de Gruyter, Berlin.

[2] Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. 2004. Understanding referring expressions involving perceptual grouping.

[3] Kotaro Funakoshi, Satoru Watanabe, and Takenobu Tokunaga. 2006, Group based generation of referring expressions

[4] Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Proceedings of the 41st Meeting of the Association for Computational Linguistics.

[5] Willem J.M. Levelt. 1989, Speaking – From Intention to Articulation. A Bradford book, The MIT Press

[6] Landragin, F., Bellalem, N. , Romary, L. (2001) Visual Salience and Perceptual Grouping in Multimodal Interactivity. In: First International Workshop on Information Presentation and Natural Multimodal Dialogue, Verona, Italy

[7] R. Moratz and T. Tenbrink. 2006. Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. Spatial Cognition and Computation.

[8] Kelleher, J. and Kruijff, G.J. (2006) Incremental Generation of Spatial Referring Expressions in Situated Dialogue. In Proceedings of COLING-ACL '06. Sydney, Australia. Association of Computational Linguistics.

[9] P. Gorniak and D. Roy. 2004. Grounded semantic composition for visual scenes. Journal of Artificial Intelligence Research, 21:429?470.

[10] K. van Deemter. 2001. Generating referring expressions: Beyond the incremental algorithm. In 4th Int. Conf. on Computational Semantics (IWCS-4), Tilburg.