A Probabilistic Model of Referring Expressions for Complex Objects

Kotaro Funakoshi[†] **Philipp Spanger**[‡] **Mikio Nakano**[†] **Takenobu Tokunaga**[‡]
[†]Honda Research Institute Japan Co., Ltd.
Saitama, Japan
funakoshi@jp.honda-ri.com
nakano@jp.honda-ri.com
[‡]Tokyo Institute of Technology
Tokyo, Japan
philipp@cl.cs.titech.ac.jp
take@cl.cs.titech.ac.jp

Abstract

This paper presents a probabilistic model both for generation and understanding of referring expressions. This model introduces the concept of *parts of objects*, modelling the necessity to deal with the characteristics of separate parts of an object in the referring process. This was ignored or implicit in previous literature. Integrating this concept into a probabilistic formulation, the model captures human characteristics of visual perception and some type of pragmatic implicature in referring expressions. Developing this kind of model is critical to deal with more complex domains in the future. As a first step in our research, we validate the model with the TUNA corpus to show that it includes conventional domain modeling as a subset.

1 Introduction

Generation of referring expressions has been studied for the last two decades. The basic orientation of this research was pursuing an algorithm that generates a minimal description which uniquely identifies a target object from distractors. Thus the research was oriented and limited by two constraints: minimality and uniqueness.

The constraint on minimality has, however, been relaxed due to the computational complexity of generation, the perceived naturalness of redundant expressions, and the easiness of understanding them (e.g., (Dale and Reiter, 1995; Spanger et al., 2008)). On the other hand, the other constraint of uniqueness has not been paid much attention to. One major aim of our research is to relax this constraint on uniqueness because of the reason explained below.

The fundamental goal of our research is to deal with multipartite objects, which have constituents with different attribute values. Typical domain settings in previous literature use uniform objects like the table A shown in Figure 1. However, real life is not so simple. Multipartite objects such as tables B and C can be found easily. Therefore this paper introduces the concept of *parts of objects* to deal with more complex domains containing such objects. Hereby the constraint on uniqueness becomes problematic because people easily generate and understand logically ambiguous expressions in such domains.

For example, people often use an expression such as "the table with red corners" to identify table B. Logically speaking, this expression is equally applicable both to A and to B, that is, violating the constraint on uniqueness. And yet people seem to have no problem identifying the intended target correctly and have little reluctance to use such an expression (Evidence is presented in Section 3). We think that this reflects some type of pragmatic implicature arising from human characteristics of visual perception and that is important both for understanding human-produced expressions and for generating human-friendly expressions in a real environment. This paper proposes a model of referring expressions both for generation and understanding. Our model uses probabilities to solve ambiguity under the relaxed constraint on uniqueness while considering human perception.

No adequate data is currently available in order to provide a comprehensive evaluation of our model. As a first step in our research, we validate the model with the TUNA corpus to show that it includes conventional domain modeling.

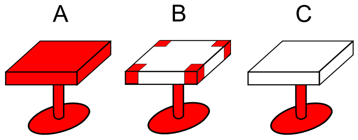

Figure 1: An example scene

Proceedings of the 12th European Workshop on Natural Language Generation, pages 191–194,
Athens, Greece, 30 – 31 March 2009. ©2009 Association for Computational Linguistics

2 Related work

Horacek (2005) proposes to introduce probabilities to overcome uncertainties due to discrepancies in knowledge and cognition between subjects. While our model shares the same awareness of issues with Horacek's work, our focus is on rather different issues (i.e., handling multipartite objects and relaxing the constraint on uniqueness). In addition, Horacek's work is concerned only with generation while our model is available both for generation and understanding. Roy (2002) also proposes a probabilistic model for generation but presupposes uniform objects.

Horacek (2006) deals with references for structured objects such as documents. Although it considers parts of objects, the motivation and focus of the work are on quite different aspects from ours.

3 Evidence against logical uniqueness

We conducted two psycholinguistic experiments using the visual stimulus shown in Figure 1.

In the first experiment, thirteen Japanese subjects were presented with an expression "kado no akai tukue (the table with red corners)" and asked to choose a table from the three in the figure. Twelve out of the thirteen chose table B. Seven out of the twelve subjects answered that the given expression was not ambiguous.

In the second experiment, thirteen different Japanese subjects were asked to make a description for table B without using positional relations. Ten out of the thirteen made expressions semantically equivalent to the expression used in the first experiment. Only three subjects made logically discriminative expressions such as "asi to yotu kado dake akai tukue (the table whose four corners and leg only are red)."

These results show that people easily generate/understand logically ambiguous expressions.

4 Proposed model

We define $\pi = \{p^1, p^2, \ldots, p^k\}$ as the set of k *parts of objects* (classes of sub-parts) that appears in a domain. Here p^1 is special and always means the whole of an object. In a furniture domain, p^1 means a piece of furniture regardless of the kind of the object (*chair*, *table*, whatever). $p^i (i \neq 1)$ means a sub-part class such as *leg*. Note that π is defined not for each object but for a domain. Thus, objects may have no part corresponding to p^i (e.g., some chairs have no leg.).

A referring expression e is represented as a set of n pairs of an attribute value expression e^a_j and a part expression e^p_j modified by e^a_j as

$$e = \{(e^p_1, e^a_1), (e^p_2, e^a_2), \ldots, (e^p_n, e^a_n)\}. \quad (1)$$

For example, an expression "the white table with a red leg" is represented as

$$\{(\text{"table"}, \text{"white"}), (\text{"leg"}, \text{"red"})\}.$$

Given a set of objects ω and a referring expression e, the probability with which the expression e refers to an object $o \in \omega$ is denoted as $Pr(O = o|E = e, \Omega = \omega)$. If we seek to provide a more realistic model, we can model a probabilistic distribution even for Ω. In this paper, however, we assume that Ω is fixed to ω and it is shared by interlocutors exactly. Thus, hereafter, $Pr(o|e)$ is equal to $Pr(o|e, \omega)$.

Following the definition (1), we estimate $Pr(o|e)$ as follows:

$$Pr(o|e) \approx \mathcal{N} \prod_i Pr(o|e^p_i, e^a_i). \quad (2)$$

Here, $\mathcal{N}$ is a normalization coefficient. According to Bayes' rule,

$$Pr(o|e^p_i, e^a_i) = \frac{Pr(o)Pr(e^p_i, e^a_i|o)}{Pr(e^p_i, e^a_i)}. \quad (3)$$

Therefore,

$$Pr(o|e) \approx \mathcal{N} \prod_i \frac{Pr(o)Pr(e^p_i, e^a_i|o)}{Pr(e^p_i, e^a_i)}. \quad (4)$$

We decompose $Pr(e^p_i, e^a_i|o)$ as

$$\sum_u \sum_v Pr(e^p_i|p_u, o)Pr(e^a_i|a_v, o)Pr(p_u, a_v|o) \quad (5)$$

where p_u is one of *parts of objects* that could be expressed with e^p_i, and a_v is one of attribute values[1] that could be expressed with e^a_i. Under the simplifying assumption that e^p_i and e^a_i are not ambiguous and are single possible expressions for a part of objects and an attribute value independently of objects [2],

$$\begin{aligned} Pr(o|e) &\approx \mathcal{N} \prod_i \frac{Pr(o)Pr(p_i, a_i|o)}{Pr(p_i, a_i)} \quad (6) \\ &\approx \mathcal{N} \prod_i Pr(o|p_i, a_i) \quad (7) \end{aligned}$$

[1] Each attribute value belongs to an attribute α, a set of attribute values. E.g., $\alpha_{color} = \{red, white, \ldots\}$.

[2] That is, we ignore *lexical selection* matters in this paper, although our model is potentially able to handle those matters including training from corpora.

$Pr(o|p, a)$ concerns *attribute selection* in generation of referring expressions. Most attribute selection algorithms presented in past work are based on set operations over multiple attributes with discrete (i.e., symbolized) values such as colors (*red, brown, white, etc*) to find a uniquely distinguishing description. The simplest estimation of $Pr(o|p, a)$ following this conventional Boolean domain modeling is

$$Pr(o|p,a) \approx \begin{cases} |\omega'|^{-1} \ (p \text{ in } o \text{ has } a) \\ 0 \ \ (p \text{ in } o \text{ does not have } a) \end{cases} \quad (8)$$

where ω' is the subset of ω, each member of which has attribute value a in its part of p.

As Horacek (2005) pointed out, however, this standard approach is problematic in a real environment because many physical attributes are non-discrete and the symbolization of these continuous attributes have uncertainties. For example, even if two objects are blue, one can be more blueish than the other. Some subjects may say it's blue but others may say it's purple. Moreover, there is the problem of logical ambiguity pointed out in Section 1. That is, even if an attribute itself is equally applicable to several objects in a logical sense, other available information (such as visual context) might influence the interpretation of a given referring expression.

Such phenomena could be captured by estimating $Pr(o|p, a)$ as

$$Pr(o|p,a) \approx \frac{Pr(a|p,o)Pr(p|o)Pr(o)}{Pr(p,a)}. \quad (9)$$

$Pr(a|p, o)$ represents the relevance of attribute value a to part p in object o. $Pr(p|o)$ represents the salience of part p in object o. The underlying idea to deal with the problem of logical ambiguity is "If some part of an object is mentioned, it should be more salient than other parts." This is related to Grice's maxims in a different way from matters discussed in (Dale and Reiter, 1995). $Pr(p|o)$ could be computed in some manner by using the saliency map (Itti et al., 1998). $Pr(o)$ is the prior probability that object o is chosen. If potential functions (such as used in (Tokunaga et al., 2005)) are used for computing $Pr(o)$, we can naturally rank objects, which are equally relevant to a given referring expression, according to distances from interlocutors.

5 Algorithms

5.1 Understanding

Understanding a referring expression e is identifying the target object $\hat{o}$ from a set of objects ω. This is formulated in a straightforward way as

$$\hat{o} = \operatorname*{argmax}_{o \in \omega} Pr(o|e). \quad (10)$$

5.2 Generation

Generation of a referring expression is choosing the best appropriate expression $\hat{e}$ to discriminate a given object $\hat{o}$ from a set of distractors. A simple formulation is

$$\hat{e} = \operatorname*{argmax}_{e \in \rho} Pr(e)Pr(\hat{o}|e). \quad (11)$$

ρ is a pre-generated set of candidate expressions for $\hat{o}$. This paper does not explain how to generate a set of candidates.

$Pr(e)$ is the generation probability of an expression e independent of objects. This probability can be learned from a corpus. In the evaluation described in Section 6, we estimate $Pr(e)$ as

$$Pr(e) \approx Pr(|e|) \prod_i Pr(\alpha_i). \quad (12)$$

Here, $Pr(|e|)$ is the distribution of expression length in terms of numbers of attributes used. $Pr(\alpha)$ is the selection probability of a specific attribute α ($SP(a)$ in (Spanger et al., 2008)).

6 Preliminary evaluation

As mentioned above, no adequate corpus is currently available in order to provide an initial validation of our model which we present in this paper. In this section, we validate our model using the TUNA corpus (the "Participant's Pack" available for download as part of the Generation Challenge 2009) to show that it includes traditional domain modeling. We use the training-part of the corpus for training our model and the development-part for evaluation.

We note that we here assume a homogeneous distribution of the probability $Pr(o|p, a)$, i.e., we are applying formula (8) here in order to calculate this probability. We first implemented our probabilistic model for the area of understanding. This means our algorithm took as input the user's selection of attribute–value pairs in the description and calculated the most likely target object. This was

Table 1: Initial evaluation of proposed model for generation in TUNA-domain

	Furniture	*People*
Total cases	80	68
Mean Dice-score	0.78	0.66

carried out for both the furniture and people domains. Overall, outside of exceptional cases (e.g., human error), our algorithm was able to distinguish the target object for all human descriptions (precision of 100%). This means it covers all the cases the original approach dealt with.

We then implemented our model for the case of generation. We measured the similarity of the output of our algorithm with the human-produced sets by using the Dice-coefficient (see (Belz and Gatt, 2007)). We evaluated this both for the Furniture and People domain. The results are summarized in Table 1.

Our focus was here to fundamentally show how our model includes traditional modelling as a subset, without much focus or effort on tuning in order to achieve a maximum Dice-score. However, we note that the Dice-score of our algorithm was comparable to the top 5-7 systems in the 2007 GRE-Challenge (see (Belz and Gatt, 2007)) and thus produced a relatively good result. This shows how our algorithm – providing a model of the referring process in a more complex domain – is applicable as well to the very simple TUNA-domain as a special case.

7 Discussion

In past work, parts of objects were ignored or implicit. In case of the TUNA corpus, while the Furniture domain ignores parts of objects, the People domain contained parts of objects such as *hair*, *glasses*, *beard*, etc. However, they were implicitly modeled by combining a pair of a part and its attribute as an attribute such as *hairColor*. One major advantage of our model is that, by explicitly modelling parts of objects, it can handle the problem of logical ambiguity that is newly reported in this paper. Although it might be possible to handle the problem by extending previously proposed algorithms in some ways, our formulation would be clearer. Moreover, our model is directly available both for generation and understanding. Referring expressions using attributes (such as discussed in this paper) and those using discourse contexts (such as "it") are separately approached in past work. Our model possibly handles both of them in a unified manner with a small extension.

This paper ignored *relations* between objects. We, however, think that it is not difficult to prepare algorithms handling relations using our model. Generation using our model is performed in a generate-and-test manner. Therefore computational complexity is a matter of concern. However, that could be controlled by limiting the numbers of attributes and parts under consideration according to relevance and salience, because our model is under the relaxed constraint of uniqueness unlike previous work.

As future work, we have to gather data to evaluate our model and to statistically train lexical selection in a new domain containing multipartite objects.

References

Anja Belz and Albert Gatt. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *Proc. the MT Summit XI Workshop Using Corpora for Natural Language Generation: Language Generation and Machine Translation (UCNLG+MT)*, pages 75–83.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

Helmut Horacek. 2005. Generating referential descriptions under conditions of uncertainty. In *Proc. ENLG 05*.

Helmut Horacek. 2006. Generating references to parts of recursively structured objects. In *Proc. ACL 06*.

L Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

Deb Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3).

Philipp Spanger, Takehiro Kurosawa, and Takenobu Tokunaga. 2008. On "redundancy" in selecting attributes for generating referring expressions. In *Proc. COLING 08*.

Takenobu Tokunaga, Tomonori Koyama, and Suguru Saito. 2005. Meaning of Japanese spatial nouns. In *Proc. the Second ACL-SIGSEM Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 93 – 100.