Modeling Referring Expressions with Bayesian Networks

Kotaro Funakoshi Mikio Nakano
Honda Research Institute Japan Co., Ltd.
8-1 Honcho, Wako,
Saitama 351-0188, Japan
{funakoshi,nakano}@jp.honda-ri.com

Takenobu Tokunaga Ryu Iida
Tokyo Institute of Technology
2-12-1 Oookayama, Meguro,
Tokyo 152-8550, Japan
{take,ryu-i}@cl.cs.titech.ac.jp

Abstract

A probabilistic approach to the resolution of referring expressions for task-oriented dialogue systems is introduced. The approach resolves descriptions (e.g., "the blue glass"), anaphora (e.g., "it"), and deixis (e.g., "this one" w/ pointing gesture) in a unified manner. In this approach, the notion of reference domains serves an important role to handle context-dependent attributes of entities and references to sets. Previously we reported the evaluation results in a puzzle solving task. This paper briefly explains the approach and discusses the issues in two work-in-progress application projects.

1 Introduction

Referring expressions (REs) can be classified into three categories: descriptions, anaphora, and deixis. Dialogue systems (DSs) are expected to handle all the three categories of REs.

We employ a Bayesian network (BN) to model a RE. One of the two major novelties of the approach is our probabilistic formulation that handles the above three kinds of REs in a unified manner. The other is bringing reference domains (RDs) (Salmon-Alt and Romary, 2001) into that formulation. RDs are sets of referents implicitly presupposed at each use of REs. By considering RDs, our approach can appropriately interpret context-dependent attributes. In addition, by treating a reference domain as a referent, REs referring to sets of entities are handled, too.

Our approach presupposes a certain amount of manual implementation of domain-dependent knowledge by developers. Therefore, it would not be suited to general text processing. However, it has the potential to be used for any task-oriented applications such as personal agents in smart phones, in-car systems, robots, etc.

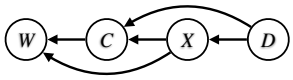

Figure 1: WCXD fundamental structure.

2 Bayesian Network-based Modeling

Each REBN (Referring Expression Bayesian Network) is tailored for a RE in the context at the moment. Its structure is determined by the syntactic and semantic information in the RE and probability tables are determined by the context. Here, we describe REBNs briefly. The details and an evaluation are found in (Funakoshi et al., 2012).

2.1 Structures

Figure 1 shows the fundamental network structure of REBNs. We call this structure WCXD. The four nodes (random variables) W, C, X, and D represent an observed word, the concept denoted by the word, the referent of the RE, and the presupposed RD, respectively. Here, a *word* means a lexical entry in the system dictionary. Each REBN is derived from the WCXD structure.

2.2 Domains of random variables

A REBN of N words referring to one entity has $2N + 2$ discrete random variables: $W_1, \ldots, W_N, C_1, \ldots, C_N, X$, and D. The domain of each variable depends on the corresponding RE and the context at the moment. Here, $\mathfrak{D}(V)$ denotes the domain of a variable V.

$\mathfrak{D}(W_i)$ contains the corresponding observed word w_i and a special symbol ω that represents other possibilities, i.e., $\mathfrak{D}(W_i) = \{w_i, \omega\}$. Each W_i has a corresponding node C_i.

$\mathfrak{D}(C_i)$ contains M concepts that can be expressed by w_i and a special concept Ω that represents other possibilities, i.e., $\mathfrak{D}(C_i) = \{c_i^1, \ldots, c_i^M, \Omega\}$. c_i^j $(j = 1 \ldots M)$ are looked up from the system dictionary.

$\mathfrak{D}(D)$ contains $L+1$ RDs recognized up to that point in time, i.e., $\mathfrak{D}(D) = \{@_0, @_1, \ldots, @_L\}$. $@_0$ is the ground domain that contains all the individual entities to be referred to in a dialogue. At the beginning of the dialogue, $\mathfrak{D}(D) = \{@_0\}$. Other L RDs are incrementally added in the course of the dialogue.

$\mathfrak{D}(X)$ contains all the possible referents, i.e., K individual entities and $L + 1$ RDs. Thus, $\mathfrak{D}(X) = \{x_1, \ldots, x_K, @_0, \ldots, @_L\}$.

2.3 Probability tables

A REBN infers the referent (i.e., the true value of node X) using four types of probability tables.

Realization model: $P(W_i|C_i, X)$

$P(W_i = w|C_i = c, X = x)$ is the probability that a hearer observes w from c and x which the speaker intends to indicate.

Relevancy model: $P(C_i|X, D)$

$P(C_i = c|X = x, D = d)$ is the probability that concept c is chosen from $\mathfrak{D}(C_i)$ to indicate x in d. Developers can implement task domain semantics in $P(C_i|X, D)$. By considering d, context-dependent attributes are handled.

Referent prediction model: $P(X|D)$

$P(X = x|D = d)$ is the probability that entity x in RD d is referred to, which is estimated according to the contextual information (such as gaze) at the time the RE is uttered but irrespective of attributive information in the RE.

Domain prediction model: $P(D)$

$P(D = d)$ is the probability that d is presupposed at the time the RE is uttered, which is estimated according to the saliency of d.

3 Work-in-Progress Apps and Issues

Currently we are working on two different applications: *Map-search* as a mobile/PC application and *Object-fetch* as a robotic application. In Map-search, the user can search locations on a map and identify a location to query the information of the location or to get a navigation to the location. In Object-fetch, the user makes a robot identify an object in the user's home or office to fetch him/her it. By applying REBNs to these domains different from each other and from the Tangram task with which we made the first evaluation, we will be able to verify the quantitative performance and qualitative ability of our approach in diverse aspects. For example, in Map-search, the number of referents can be huge while Tangram has only 7 referents. Therefore, computational complexity will be an important issue for realtime operation. It is unrealistic to consider all locations every time. We will have to devise a way to efficiently limit the number of candidates for each time without excluding true referents.

Not limited to Object-fetch but especially in it, handling of unknown objects is vital, while all objects are known in Tangram. The robot must recognize a RE to an object that it does not know. For this purpose we can introduce χ for an unknown referent in $\mathfrak{D}(X)$. Hopefully, χ will have the highest probability for REs to unknown objects. Uncertainty due to speech recognition errors, unknown words, and unknown concepts is also a severe issue. There is a possibility that adjusting the parameter ϵ (here, $P(W = w|C = \Omega, X) = \epsilon$) eases the problem. The larger ϵ is, the more $P(X|D)$ influences inference results, i.e., the contextual information outside the RE gets more importance. For example, in a low signal-noise ratio environment, the robot could selectively rely on the context by increasing ϵ.

In both applications, spatial relations are important to identify referents. To handle relations, we are going to introduce another type of node for relations in REBNs to combine multiple REBNs into one.

System design methodology is the last but not least issue. While REBNs allow different design patterns of the world inherent in each application, the best design pattern seems to depend on each. For example, using the set of the location IDs in a database as $\mathfrak{D}(X)$ seems reasonable for Map-search. However, this design pattern does not work with Object-fetch. Object-fetch requires the object IDs in the robot's database to be included in $\mathfrak{D}(C)$. Through building Map-search and Object-fetch in parallel, we would like to clarify different design patterns and the conditions to chose a design pattern for each application.

References

K Funakoshi, M. Nakano, T. Tokunaga, and R. Iida. 2012. A unified probabilistic approach to referring expressions. In *Proc. SIGDIAL 2012*.

S. Salmon-Alt and L. Romary. 2001. Reference resolution within the framework of cognitive grammar. In *Proc. Intl. Colloqium on Cognitive Science*.