A Unified Probabilistic Approach to Referring Expressions

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 Ookayama, Meguro,
Tokyo 152-8550, Japan
{take,ryu-i}@cl.cs.titech.ac.jp

Abstract

This paper proposes a probabilistic approach to the resolution of referring expressions for task-oriented dialogue systems. The approach resolves descriptions, anaphora, and deixis 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. The evaluation with the REX-J corpus shows promising results.

1 Introduction

Referring expressions (REs) are expressions intended by speakers to identify entities to hearers. REs can be classified into three categories: descriptions, anaphora, and deixis; and, in most cases, have been studied within each category and with a narrowly focused interest. Descriptive expressions (such as "the blue glass on the table") exploit attributes of entities and relations between them to distinguish an entity from the rest. They are well studied in natural language generation, e.g., (Dale and Reiter, 1995; Krahmer et al., 2003; Dale and Viethen, 2009). Anaphoric expressions (such as "it") refer to entities or concepts introduced in the preceding discourse and are studied mostly on textual monologues, e.g., (Kamp and Reyle, 1993; Mitkov, 2002; Ng, 2010). Deictic (exophoric) expressions (such as "this one") refer to entities outside the preceding discourse. They are often studied focusing on pronouns accompanied with pointing gestures in physical spaces, e.g., (Gieselmann, 2004).

Dialogue systems (DSs) as natural human-machine (HM) interfaces are expected to handle all the three categories of referring expressions (Salmon-Alt and Romary, 2001). In fact, the three categories are not mutually exclusive. To be concrete, a descriptive expression in conversation is either deictic or anaphoric. It is, however, not easy to tell whether a RE is deictic or anaphoric in advance of a resolution (regardless of whether the RE is descriptive or not). Therefore, we propose a general unified approach to the above three kinds of REs.

We employ a Bayesian network (BN) to model a RE. Dealing with continuous information and vague situations is critical to handle real world problems. Probabilistic approaches enable this for reference resolvers. Each BN is dynamically constructed based on the structural analysis result of a RE and contextual information available at that moment. The BN is used to estimate the probability with which the corresponding RE refers to an entity.

One of the two major contributions of this paper is our probabilistic formulation that handles the above three kinds of REs in a unified manner. Previously Iida et al. (2010) proposed a quantitative approach that handles anaphoric and deictic expressions in a unified manner. However it lacks handling of descriptive expressions. Our formulation subsumes and extends it to handle descriptive REs. So far, no previously proposed method for reference resolution handles all three types of REs.

The other contribution is bringing *reference domains* into that formulation. Reference domains (Salmon-Alt and Romary, 2000) are sets of referents implicitly presupposed at each use of REs. By considering them, 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. As far as the authors know, this work is the first that takes a probabilistic approach to reference domains.

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 237–246,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

1.1 Reference domains

First, we explain reference domains concretely. Reference domains (RDs) (Salmon-Alt and Romary, 2000; Salmon-Alt and Romary, 2001; Denis, 2010) are theoretical constructs, which are basically sets of entities presupposed at each use of REs. RDs in the original literature are not mere sets of entities but mental objects equipped with properties such as *type, focus, or saliency* and internally structured with *partitions*. In this paper, while we do not explicitly handle partitions, reference domains can be nested as an approximation of partitioning, that is, an entity included in a RD is either an individual entity or another RD. Each RD d has its focus and degree of saliency (a non-negative real number). Hereafter, two of them are denoted as $\mathrm{foc}(d)$ and $\mathrm{sal}(d)$ respectively. RDs are sorted in descending order according to saliency.

We illustrate reference domains with figure 1. It shows a snapshot of solving a Tangram puzzle (the puzzle and corpus are explained in section 3.1). RDs are introduced into our mental spaces either linguistically (by hearing a RE) or visually (by observing a physical situation). If one says "the two big triangles" in the situation shown in figure 1, we will recognize a RD consisting of pieces 1 and 2. If we observe one moves piece 1 and attaches it to piece 2, we will perceptually recognize a RD consisting of pieces 1, 2, and 6 due to proximity (Thórisson, 1994). In a similar way, a RD consisting of pieces 5 and 7 also can be recognized. Hereafter, we indicate a RD with the mark @ with an index, and denote its elements by enclosing them with [ ]. E.g., $@_1 = [1, 2]$, $@_2 = [1, 2, 6]$, $@_3 = [5, 7]$. The focused entity is marked by '*'. Thus, $\mathrm{foc}([1*, 2]) = 1$.

The referent of a RE depends on which RD is presupposed. That is, if one presupposes $@_1$ or $@_2$, the referent of "the right piece" should be piece 1. If one presupposes $@_3$, the referent of the same RE should be piece 5. This is the context-dependency mentioned above.

Previous work on RDs (Salmon-Alt and Romary, 2000; Salmon-Alt and Romary, 2001; Denis, 2010) employ not probabilistic but formal approaches.

1.2 Probabilistic approaches to REs

Here, previous probabilistic approaches to REs are explained and differences between ours and theirs are highlighted. Bayesian networks (Pearl, 1988; Jensen and Nielsen, 2007) have been not often but occasionally applied to problems in natural language processing/computational linguistics since (Charniak and Goldman, 1989). With regard to REs, Burger and Connolly (1992) proposed a BN specialized for anaphora resolution. Weissenbacher (2005; 2007) proposed a BN for the resolution of non-anaphoric "it" and also a BN for the resolution of pronominal anaphora. They used pre-defined fixed BNs for their tasks while our approach dynamically tailors a BN for each RE.

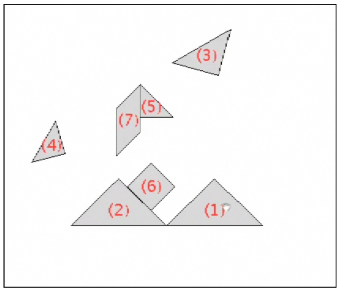

Figure 1: Tangram puzzle. (The labels 1 to 7 are for illustration purposes and not visible to participants.)

Cho and Maida (1992) and Roy (2002) adopted not exactly BNs but similar probabilistic approaches for reference resolution and generation respectively. However, their foci are only on descriptions.

Lison et al. (2010) proposed an approach using Markov logic networks (MLNs) (Richardson and Domingos, 2006) to reference resolution. They dealt with only deictic and descriptive REs. Even though MLNs are also a probabilistic framework, it is difficult for DS developers to provide quantitative domain knowledge needed to resolve REs because MLNs accept domain knowledge in the form of formal logic rules with weights, which must be determined globally. In contrast, BNs are more flexible and easy in providing quantitative knowledge to DSs in the form of conditional probability tables, which can be determined locally.

As just described, there are several probabilistic approaches to REs but none of them incorporates reference domains. In the next section, we introduce our *REBNs (Referring Expression Bayesian Networks)*, a novel Bayesian network-based modeling approach to REs that incorporates reference domains.

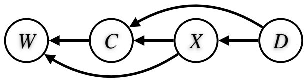

Figure 2: WCXD fundamental structure.

2 Bayesian Network-based Modeling of Referring Expressions

Each REBN is dedicated 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.

2.1 Structures

Figure 2 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 defined by the DS developer (concept dictionary; section 3.2.1).

Each REBN is constructed by modifying or multiply connecting the WCXD structure as shown in figures 3 and 4. Figure 3 shows the network for REs indicating one referent such as "that table." Each W_i node has a corresponding word w_i. Figure 4 shows the network for REs indicating two referents such as "his table." We call the class of the former REs *s-REX* (simple Referring EXpression) and the class of the latter REs *c-REX* (compound Referring EXpression). Although REBNs have the potential to deal with c-REX, hereafter we concentrate on s-REX because the page space is limited and the corpus used for evaluation contains very few c-REX instances.

Although, in section 1, we explained that (Iida et al., 2010) handles anaphoric and deictic expressions in a unified manner, it handles anaphora to instances only and does not handle that to concepts. Therefore, it cannot satisfactorily resolve such an expression "Bring me the red box, and the blue one, too." Here, "one" does not refer to the physical referent of "the red box" but refers to the concept of "box". The C nodes will enable handling of such references to concepts. This is one of the important features of REBNs but will be investigated in future work.

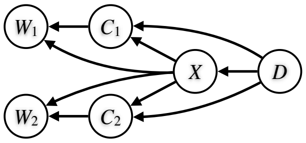

Figure 3: BN for two-word REs indicating one referent.

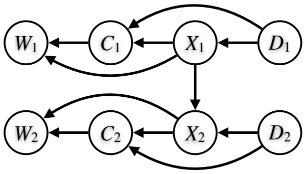

Figure 4: BN for two-word REs indicating two referents.

2.2 Domains of random variables

A REBN for an s-REX instance of N words 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 random 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 concept dictionary (see section 3.2.1, table 2).

$\mathfrak{D}(D)$ 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, $\mathfrak{D}(X) = \{x_1, \ldots, x_K, @_0, \ldots, @_L\}$. Including RDs enables handling of references to sets.

Then reference resolution is formalized as below:

$$x' = \underset{x \in \mathfrak{D}(X)}{\arg\max}\, P(X = x | W_1 = w_1, \ldots, W_N = w_N). \quad (1)$$

$P(X|W_1, \ldots, W_N)$ is obtained by marginalizing the joint probabilities that are computed with the probability tables described in the next subsection.

2.3 Probability tables

Probability distributions are given as (conditional) probability tables since all the random variables used in a REBN are discrete. Here, four types of probability tables used by REBNs are described.

2.3.1 $P(W_i|C_i, X)$

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

In most cases, W_i does not depend on X, i.e., $P(W_i|C_i, X) \equiv P(W_i|C_i)$. X is, however, necessary to handle individualized terms (names).

There are several conceivable ways of probability assignment. One simple way is: for each c_i^j, $P(W = w_i|C = c_i^j) = 1/T, P(W = \omega|C = c_i^j) = (T-1)/T$, and for Ω, $P(W = w_i|C = \Omega) = \epsilon, P(W = \omega|C = \Omega) = 1 - \epsilon$. Here T is the number of possible words for c_i^j. ϵ is a predefined small number such as 10^{-8}. We use this assignment in the evaluation.

2.3.2 $P(C_i|X, D)$

$P(C_i = c|X = x, D = d)$ is the probability that concept c is chosen from $\mathfrak{D}(C_i)$ to indicate x in d.

The developers of DSs cannot provide $P(C_i|X, D)$ in advance because $\mathfrak{D}(C_i)$ is context-dependent. Therefore, we take an approach of composing $P(C_i|X = x, D = d)$ from $R(c_i^j, x, d)$ ($c_i^j \in \mathfrak{D}(C_i)\backslash\{\Omega\}$). Here $R(c_i^j, x, d)$ is the relevancy of concept c_i^j to referent x with regard to d, and $0 \leq R(c_i^j, x, d) \leq 1$. 1 means full relevancy and 0 means no relevancy. 0.5 means neutral. For example, a concept BOX will have a high relevancy to a suitcase such as 0.8 but a concept BALL will have a low relevancy to the suitcase such as 0.1. If x is not in d, $R(c_i^j, x, d)$ is 0. Algorithm 1 in appendix A shows an algorithm to compose $P(C_i|X = x, D = d)$ from $R(c_i^j, x, d)$. Concept Ω will be assigned a high probability if none of $c_i^j \in \mathfrak{D}(C_i)\backslash\{\Omega\}$ has a high relevancy to x.

If c_i^j is static,[1] $R(c_i^j, x, d)$ is numerically given in advance in the form of a table. If not static, it is implemented as a function by the DS developer, that is, $R(c_i^j, x, d) = f_{c_i^j}(x, d, I)$. Here I is all the information available from the DS.

For example, given a situation such as shown in figure 1, the relevancy function of a positional concept LEFT (suppose a RE such as "the left piece") can be implemented as below:

$$f_{\text{LEFT}}(x, d, I) = (u_x - u_r)/(u_l - u_r). \qquad (2)$$

Here, u_x, u_l and u_r are respectively the horizontal coordinates of x, the leftmost piece in d, and the rightmost piece in d, which are obtained from I. If x is a RD, the relevancy is given as the average of entities included in the RD.

2.3.3 $P(X|D)$

$P(X = x|D = d)$ is the probability that entity x in RD d is referred to, which is estimated according to the contextual information at the time the corresponding RE is uttered but irrespective of attributive information in the RE. The contextual information includes the history of referring so far (discourse) and physical statuses such as the gaze of the referrer (situation). We call $P(X = x|D = d)$ the *prediction model.*

The prediction model can be constructed by using a machine learning-based method. We use a ranking-based method (Iida et al., 2010). The score output by the method is input into the standard sigmoid function and normalized to be a probability. If x is not in d, $P(X = x|D = d)$ is 0.

2.3.4 $P(D)$

$P(D = d)$ is the probability that RD d is presupposed at the time the RE is uttered. We cannot collect data to estimate this probabilistic model because RDs are implicit. Therefore, we examine three a priori approximation functions based on the saliency of d. Saliency is proportional to recency.[2]

Uniform model This model ignores saliency. This is introduced to see the importance of saliency.

$$P(D = d) = 1/|\mathfrak{D}(D)| \qquad (3)$$

Linear model This model distributes probabilities in proportion to saliency. This is an analogy of the method used in (Denis, 2010).

$$P(D = d) = \frac{\text{sal}(d)}{\sum_{d' \in \mathfrak{D}(D)} \text{sal}(d')} \qquad (4)$$

[1] Whether a concept is static or not depends on each DS.

[2] Assignment of saliency is described in section 3.2.3.

Exponential model This model puts emphasis on recent RDs. This function is so called soft-max.

$$P(D = d) = \frac{\exp(\text{sal}(d))}{\sum_{d' \in \mathfrak{D}(D)} \exp(\text{sal}(d'))} \quad (5)$$

3 Experimental Evaluation

We evaluated the potential of the proposed framework by using a situated human-human (HH) dialogue corpus.

3.1 Corpus

We used the REX-J Japanese referring expression corpus (Spanger et al., 2010). The REX-J corpus consists of 24 HH dialogues in each of which two participants solve a Tangram puzzle of seven pieces (see figure 1). The goal of the puzzle is combining seven pieces to form a designated shape (such as a swan). One of two subjects takes the role of operator (OP) and the other takes the role of solver (SV). The OP can manipulate the virtual puzzle pieces displayed on a PC monitor by using a computer mouse but does not know the goal shape. The SV knows the goal shape but cannot manipulate the pieces. The states of the pieces and the mouse cursor operated by the OP are shared by the two subjects in real time. Thus, the two participants weave a collaborative dialogue including many REs to the pieces. In addition to REs, the positions and directions of the pieces, the position of the mouse cursor, and the manipulation by the OP were recorded with timestamps and the IDs of relevant pieces.

3.1.1 Annotation

Each RE is annotated with its referent(s) as shown in table 1. The 1st RE *okkiisankaku*[3] big triangle "a big triangle" in the table is ambiguous and refers to either piece 1 or 2. The 7th and 8th REs refer to the set of pieces 1 and 2. The other REs refer to an individual piece.

To skip the structural analysis of REs to avoid problems due to errors in such analysis, we have additionally annotated the corpus with intermediate structures, from which REBNs are constructed. Because we focus on s-REX only in this paper, the intermediate structures are straightforward:[4] parenthesized lists of separated words as shown in table 1. The procedure to generate a REBN of s-REX from such an intermediate structure is also straightforward and thus it is not explained due to the page limitation.

3.2 Implementations

We use BNJ[5] for probabilistic computation. Here we describe the implementations of resources and procedures that are more or less specific to the task domain of REX-J.

3.2.1 Concept dictionary

Table 2 shows an excerpt of the concept dictionary defined for REX-J. We manually defined 40 concepts by observing the dialogues.

3.2.2 Static relevancy table and relevancy functions

For 13 concepts out of 40, their relevancy values were manually determined by the authors. Table 3 shows an excerpt of the static relevancy table defined for the seven pieces shown in figure 1. TRI is relevant only to pieces 1 to 5, and SQR is relevant only to pieces 6 and 7 but is not totally relevant to piece 7 because it is not a square in a precise sense. FIG is equally but not very relevant to all the pieces,[6]

For the remaining 27 concepts, we implemented relevancy functions (see appendix B).

3.2.3 Updating the list of RDs

In our experiment, REs are sequentially resolved from the beginning of each dialogue in the corpus. In the course of resolution, RDs are added into a list and updated by the following procedure. RDs are sorted in descending order according to saliency.

At each time of resolution, we assume that all the previous REs are correctly resolved. Therefore, after each time of resolution, if the correct referent of the last RE is a set, we add a new RD equivalent to the set into the list of RDs, unless the list contains another equivalent RD already. In either case, the saliency of the RD equivalent to the set is set to $\sigma + 1$ unless the RD is at the head of the list already.

[3] Words are not separated by white spaces in Japanese.

[4] In the case of c-REX, graph-like structures are required.

[5] http://bnj.sourceforge.net/

[6] This is because concept FIG in REX-J is usually used to refer to not a single piece but a shaped form (combined pieces).

D-ID	Role	Start	End	Referring expression	Referents	Intermediate structure
0801	SV	17.345	18.390	*okkiisankaku* big triangle	1 or 2	(*okkii sankaku*)
0801	SV	20.758	21.368	*sore* it	1	(*sore*)
0801	SV	23.394	24.720	*migigawanookkiisankaku* right big triangle	1	(*migigawano okkii sankaku*)
0801	SV	25.084	25.277	*kore* this	1	(*kore*)
0801	SV	26.512	26.671	*sono* that	1	(*sono*)
0801	SV	28.871	29.747	*konookkiisankaku* this big triangle	2	(*kono okkii sankaku*)
0801	OP	46.497	48.204	*okkinasankakkei* big triangle	1, 2	(*okkina sankakkei*)
0801	OP	51.958	52.228	*ryôhô* both	1, 2	(*ryôhô*)

“D-ID” means dialogue ID. “Start” and “End” mean the end points of a RE.

Table 1: Excerpt of the corpus annotation (w/ English literal translations).

Concept	Words
TRI	triangle, right triangle
SQR	quadrate, square, regular tetragon
FIG	figure, shape

Table 2: Dictionary (excerpted and translated in English).

Concept	Relevancy values by piece						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
TRI	1	1	1	1	1	0	0
SQR	0	0	0	0	0	1	0.8
FIG	0.3	0.3	0.3	0.3	0.3	0.3	0.3

Table 3: Static relevancy table.

Here, σ is the largest saliency value in the list at the moment (the saliency value of the head RD).

Before each time of resolution, we check whether the piece that is most recently manipulated after the previous RE constitutes a perceptual group by using the method explained in section 3.2.4 at the onset time of the target RE. If such a group is recognized, we add a new RD equivalent to the recognized group unless the list contains another equivalent RD. In either case, the saliency of the RD equivalent is set to $\sigma + 1$ unless the RD is at the head of the list already, and the focus of the equivalent RD is set to the most recently manipulated piece.

When a new RD $@_m$ is added to the list, a complementary RD $@_n$ and a subsuming RD $@_l$ are also inserted just after $@_m$ in the list. Here, $@_n = @_0 \backslash @_m$ and $@_l = [@_m *, @_n]$. This operation is required to handle a concept REST, e.g., “the remaining pieces.”

3.2.4 Perceptual grouping

There is a generally available method of simulated perceptual grouping (Thórisson, 1994). It works well in a spread situation such as shown in figure 1 but tends to produce results that do not match our intuition when pieces are tightly packed at the end of a dialogue. Therefore, we adopt a simple method that recognizes a group when a piece is attached to another. This method is less general but works satisfactorily in the REX-J domain due to the nature of the Tangram puzzle.

3.2.5 Ranking-based prediction model

As mentioned in section 2.3.3, a ranking-based method (Iida et al., 2010) using SVM^{rank} (Joachims, 2006) was adopted for constructing the prediction model $P(X|D)$. This model ranks entities according to 16 binary features such as *whether the target entity is previously referred to* (a discourse feature), *whether the target is under the mouse cursor* (a mouse cursor feature), etc.[7]

When a target is a set (i.e., a RD), discourse features for it are computed as in the case of a piece; meanwhile, mouse cursor features are handled in a different manner. That is, if one of the group members meets the criterion of a mouse cursor feature, the group is judged as meeting the criterion.

In (Iida et al., 2010), preparing different models for pronouns and non-pronouns achieved better performance. Therefore we trained two linear kernel SVM models for pronouns and non-pronouns with the 24 dialogues.

3.3 Experiment

We used the 24 dialogues for evaluation.[8] As mentioned in section 2.1, we focused on s-REX. These 24 dialogues contain 1,474 s-REX instances and 28 c-REX instances. In addition to c-REX, we excluded REs mentioning complicated concepts, for which it is difficult to implement relevancy functions in a short time.[9] After excluding those REs,

[7] Following the results shown in (Iida et al., 2010), we did not use the 6 manipulation-related features (CO1 ... CO6).

[8] We used the same data to train the SVM-rank models. This is equivalent to assuming that we have data large enough to saturate the performance of the prediction model.

[9] Mostly, those are metaphors such as “neck” and concepts related to operations such as “put.” For example, although

$P(D)$ model	Most-recent			Mono-domain			Uniform			Linear			Exponential		
Category	Single	Plural	Total	Single	Plural	Total	Single	Plural	Total	Single	Plural	Total	Single	Plural	Total
w/o S/P info.	42.4	28.8	40.0	77.5	47.3	73.3	77.1	40.6	72.0	78.3	45.1	73.7	76.2	48.4	72.3
w/ S/P info.	44.3	35.4	42.7	84.8	58.8	81.2	84.4	55.0	80.3	85.6	61.0	82.1	83.4	68.1	81.3

Table 4: Results of reference resolution (Accuracy in %).

1,310 REs were available. Out of the 1,310 REs, 182 REs (13.9%) refers to sets, and 612 REs (46.7%) are demonstrative pronouns such as *sore* "it."

3.3.1 Settings

We presupposed the following conditions.

Speaker role independence: We assumed REs are independent of speaker roles, i.e., SV and OP. All REs were mixed and processed serially.

Perfect preprocessing and past information: As mentioned in sections 3.1.1 and 3.2.3, we assumed that no error comes from preprocessing including speech recognition, morphological analysis, and syntactic analysis;[10] and all the correct referents of past REs are known.[11]

No future information: In HH dialogue, sometimes information helpful for resolving a RE is provided after the RE is uttered. We, however, do not consider such future information.

Numeral information: Many languages including English grammatically require indication of numeral distinctions by using such as articles, singular/plural forms of nouns and copulas, etc. Although Japanese does not have such grammatical devices,[12] it would be possible to predict such distinctions by using a machine learning technique with linguistic and gestural information. Therefore we observed the effect of providing such information. In the following experiment we provide the singular/plural distinction information to REBNs by looking at the annotations of the correct referents in advance. This is achieved by adding a special evidence node C_0, where $\mathfrak{D}(C_0) = \{\mathrm{S}, \mathrm{P}\}$. $P(C_0 = \mathrm{S}|X = x) = 1$ and $P(\mathrm{P}|x) = 0$ if x is a piece. On the contrary, $P(\mathrm{S}|x) = 0$ and $P(\mathrm{P}|x) = 1$ if x is a set.

3.3.2 Baselines

To our best knowledge, there is no directly comparable method. We set up two baselines. The first baseline uses the most recent as the resolved referent for each RE (Initial resolution of each dialogue always fails). This baseline is called *Most-recent*.

As the second baseline, we prepared another $P(D)$ model in addition to those explained in section 2.3.4, which is called *Mono-domain*. In Mono-domain, $\mathfrak{D}(D)$ consists of only a single RD $@'_0$, which contains individual pieces and the RDs recognized up to that point in time. That is, $@'_0 = \mathfrak{D}(X)$. Resolution using this model can be considered as a straightforward extension of (Iida et al., 2010), which enables handling of richer concepts in REs[13] and handling of REs to sets[14].

3.3.3 Results

The performance of reference resolution is presented by category and by condition in terms of accuracy (# of correctly resolved REs/# of REs).

We set up the three categories in evaluating resolution, that is, Single, Plural, and Total. Category Single is the collection of REs referring to a single piece. Plural is the collection of REs referring to a set of pieces. Total is the sum of them. Ambiguous REs such as the first one in table 1 are counted as "Single" and the resolution of such a RE is considered correct if the resolved result is one of the possible referents.

"putting a piece" and "getting a piece out" are distinguished due to speakers' intentions, they are (at least superficially) homogeneous in the physical data available from the corpus and difficult for machines to distinguish each other.

[10] In general, the speech and expressions in human-machine (HM) dialogue are less complex and less difficult to process than those in HH dialogue data. This is typcially observed as fewer disfluencies (Shriberg, 2001) and simpler sentences with fewer omissions (Itoh et al., 2002). Therefore, when we apply our framework to real DSs, we can expect clearer and simpler input and thus better performance. We supposed that the condition of perfect preprocessing in HH dialogue approximates the results to those obtained when HM dialogue data is used.

[11] If a reference is misinterpreted (i.e., wrongly resolved) in a dialogue, usually that misinterpretation will be repaired by the interlocutors in the succeeding interaction once the misinterpretation becomes apparent. Therefore, accumulating all past errors in resolution is rather irrational as an experimental setting.

[12] Japanese has a plurality marker *-ra* (e.g., *sore-ra*), but use of it is not mandatory (except for personal pronouns).

[13] (Iida et al., 2010) used only object types and sizes. Other concepts such as LEFT were simply ignored.

[14] (Iida et al., 2010) did not deal with REs to sets.

“w/o S/P info.” indicates experimental results without singular/plural distinction information. “w/ S/P info.” indicates experimental results with it.

Table 4 shows the results of reference resolution per $P(D)$ modeling method.[15] Obviously S/P information has a significant impact.

While the best performance for category Single was achieved with the Linear model, the best performance for Plural was achieved with the Exponential model. If it is possible to know whether a RE is of Single or Plural, that is, if S/P information is available, we can choose a suitable $P(D)$ model. Therefore, by switching models, the best performance of Total with S/P information reached 83.4%, and a gain of 2.0 points against Mono-domain was achieved (sign test, $p < 0.0001$).

Because the corpus did not include many instances to which the notion of reference domains is effective, the impact of RDs may appear small on the whole. In fact, the impact was not small. By introducing RDs, resolution in category Plural achieved a significant advancement. The highest gain from Mono-domain was 9.3 points (sign test, $p < 0.005$). Moreover, more REs containing positional concepts such as LEFT and RIGHT were correctly resolved in the cases of Uniform, Linear, and Exponential. Table 5 summarizes the resolution results of four positional concepts (with S/P information). While Mono-domain resolved 65% of them, Linear correctly resolved 75% (sign test, $p < 0.05$).

As shown in table 4, the performance of the Uniform model was worse than that of Mono-domain. This indicates that RDs introduced without an appropriate management of them would be harmful noise. Conversely, it also suggests that there might be a room for improvement by looking deeply into the management of RDs (e.g., *forgetting* old RDs).

4 Conclusion

This paper proposed a probabilistic approach to reference resolution, REBNs, which stands for Referring Expression Bayesian Networks. At each time of resolution, a dedicated BN is constructed for the RE in question. The constructed BN deals with either descriptive, deictic or anaphoric REs in a unified manner. REBNs incorporate the notion of reference domains (RDs), which enables the resolution of REs with context-dependent attributes and handling of REs to sets. REBNs are for task-oriented dialogue systems and presuppose a certain amount of domain-dependent manual implementation by developers. Therefore, REBNs would not be suited to general text processing or non-task-oriented systems. However, REBNs have the potential to be a standard approach that can be used for any and all task-oriented applications such as personal agents in smart phones, in-car systems, service robots, etc.

Concept	Count	Mono	Uni.	Lin.	Exp.
LEFT	21	11	12	16	13
RIGHT	33	23	23	25	27
UPPER	9	6	6	6	4
LOWER	6	5	4	5	4
Total	69	45	45	52	48

(*Count means the numbers of occurrence of each concept. Mono, Uni., Lin., and Exp. correspond to Mono-domain, Uniform, Linear and Exponential.*)

Table 5: Numbers of correctly resolved REs containing positional concepts.

The proposed approach was evaluated with the REX-J human-human dialogue corpus and promising results were obtained. The impact of incorporating RDs in the domain of the REX-J corpus was recognizable but not so large on the whole. However, in other types of task domains where grouping and comparisons of objects occur frequently, the impact would be larger. Note that REBNs are not limited to Japanese, even though the evaluation used a Japanese corpus. Evaluations with human-machine dialogue are important future work.

Although this paper focused on the simple type of REs without relations, REBNs are potentially able to deal with complex REs with relations. The evaluation for complex REs is necessary to validate this potential of REBN. Currently REBN assumes REs whose referents are concrete entities. An extension for handling abstract entities (Byron, 2002; Müller, 2007) is important future work. Another direction would be generating REs with REBNs. A generate-and-test approach is a naive application of REBN for generation. More efficient method is, however, necessary.

[15] According to the results of preliminary experiments, even in the case of the Uniform/Linear/Exponential models, we resolved the REs having demonstratives with the Mono-domain model. This is in line with the finding of separating models between pronouns and non-pronouns in (Iida et al., 2010).

References

John D. Burger and Dennis Connoly. 1992. Probabilistic resolution of anaphoric reference. In *Proceedings of the AAAI Fall Symposium on Intelligent Probabilistic Approaches to Natural Language*, pages 17–24.

Donna Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 80–87.

Eugene Charniak and Robert Goldman. 1989. A semantics for probabilistic quantifier-free first-order languages with particular application to story understanding. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1074–1079, Menlo Park, CA, USA.

Sehyeong Cho and Anthony Maida. 1992. Using a Bayesian framework to identify the referent of definite descriptions. In *Proceedings of the AAAI Fall Symposium on Intelligent Probabilistic Approaches to Natural Language*, pages 39–46.

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

Robert Dale and Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of the the 12th European Workshop on Natural Language Generation (ENLG)*, pages 59–65, Athens, Greece, March.

Alexandre Denis. 2010. Generating referring expressions with reference domain theory. In *Proceedings of the 6th International Natural Language Generation Conference (INLG)*, pages 27–35.

Petra Gieselmann. 2004. Reference resolution mechanisms in dialogue management. In *Proceedings of the 8th workshop on the semantics and pragmatics of dialogue (CATALOG)*, pages 28–34, Barcelona, Italy, July.

Ryu Iida, Shumpei Kobayashi, and Takenobu Tokunaga. 2010. Incorporating extra-linguistic information into reference resolution in collaborative task dialogue. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1259–1267, Uppsala, Sweden, July.

Toshihiko Itoh, Atsuhiko Kai, Tatsuhiro Konishi, and Yukihiro Itoh. 2002. Linguistic and acoustic changes of user's utterances caused by different dialogue situations. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, pages 545–548.

Finn V. Jensen and Thomas D. Nielsen. 2007. *Bayesian Networks and Decision Graphs*. Springer, second edition.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226, Philadelphia, PA, USA, August.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers.

Emiel Krahmer, Sebastiaan van Erk, and André Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29:53–72.

Pierre Lison, Carsten Ehrler, and Geert-Jan M. Kruijff. 2010. Belief modelling for situation awareness in human-robot interaction. In *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 138–143, Viareggio, Italy, September.

Ruslan Mitkov. 2002. *Anaphora Resolution*. Studies in Language and Linguistics. Pearson Education.

Christoph Müller. 2007. Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 816–823.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July.

Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, USA.

Matthew Richardson and Pedor Domingos. 2006. Markov logic networks. *Machine Learning*, 62(1–2):107–136.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3):353–385.

Susanne Salmon-Alt and Laurent Romary. 2000. Generating referring expressions in multimodal contexts. In *Proceedings of the INLG 2000 workshop on Coherence in Generated Multimedia*, Mitzpe Ramon, Israel, June.

Susanne Salmon-Alt and Laurent Romary. 2001. Reference resolution within the framework of cognitive grammar. In *Proceedings of the International Colloqium on Cognitive Science*, San Sebastian, Spain, May.

Elizabeth Shriberg. 2001. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169.

Philipp Spanger, Masaaki Yasuhara, Ryu Iida, Takenobu Tokunaga, Asuka Terai, and Naoko Kuriyama. 2010. REX-J: Japanese referring expression corpus of situated dialogs. *Language Resources and Evaluation*. Online First, DOI: 10.1007/s10579-010-9134-8.

Kristinn R. Thórisson. 1994. Simulated perceptual grouping: An application to human-computer interaction. In *Proceedings of the 16th Annual Conference of the Cognitive Science Society*, pages 876–881, Atlanta, GA, USA.

Davy Weissenbacher and Adeline Nazarenko. 2007. A Bayesian approach combining surface clues and linguistic knowledge: Application to the anaphora resolution problem. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Davy Weissenbacher. 2005. A Bayesian network for the resolution of non-anaphoric pronoun it. In *Proceedings of the NIPS 2005 Workshop on Bayesian Methods for Natural Language Processing*, Whistler, BC, Canada.

A Algorithm to compose $P(C|X, D)$

Algorithm 1 Composing $P(C|X = x, D = d)$.

```
Input: $\mathfrak{D}(C)$; $R(c, x, d)$ for all $c \in \mathfrak{D}(C) \backslash \{\Omega\}$
Output: $P(C|X = x, D = d)$
$n \leftarrow 0, s \leftarrow 0, S = \mathfrak{D}(C) \backslash \{\Omega\}$
for all $c \in S$ do
    $r[c] \leftarrow R(c, x, d)$ #{Relevancy of concept $c$}
    $s \leftarrow s + r[c]$ #{Sum of relevancy $r[c]$}
    $n \leftarrow n + (1 - r[c])$ #{Sum of residual $(1 - r[c])$}
end for
$r[\Omega] \leftarrow n/|S|$
$s \leftarrow s + r[\Omega]$
for all $c \in \mathfrak{D}(C)$ do
    $P(C = c|X = x, D = d) \leftarrow r[c]/s$
end for
```

(#{...} is a comment.)

B Relevancy functions

As explained in section 2.3.2, the relevancy functions for positional concepts such as LEFT and RIGHT were implemented as geometric calculations. Here several other relevancy functions are shown with corresponding example REs.

"this *figure*":

$$R(\text{FIG}, x, d) = \begin{cases} 0.3 & : \text{ if single(x)} \\ 1 & : \text{ if not single}(x) \text{ and shape}(x) \\ 0 & : \text{ otherwise} \end{cases}$$

(single(x) means x is a single piece. shape(x) means x is a set of pieces that are concatenated and form a shape. 0.3 comes from the static relevancy table.)

"*both* the triangles":

$$R(\text{BOTH}, x, d) = \begin{cases} 1 & : \text{ if } |x| = 2 \\ 0 & : \text{ otherwise} \end{cases}$$

"*another* one":

$$R(\text{ANOTHER}, x, d) = \begin{cases} 1 & : \text{ if foc}(d) \neq x \\ 0 & : \text{ otherwise} \end{cases}$$

"the *remaining* ones":

$$R(\text{REST}, x, d) = \begin{cases} 1 & : \text{ if } d = [x, y*] \\ 0 & : \text{ otherwise} \end{cases}$$

(REST requires $|d| = 2$, and both x and y are sets. ANOTHER does not.)

"*all*":

$$R(\text{ALL}, x, d) = \begin{cases} 1 & : \text{ if } x = d \\ 0 & : \text{ otherwise} \end{cases}$$

(ALL does not always refer to $@_0$.)