不均一な物体を考慮した参照表現の確率的モデル

A Probabilistic Model of Referring Expressions for Complex Objects

船越 孝太郎[*1] スパンガー フィリップ[*2] 中野 幹生[*1] 徳永 健伸[*2]
Kotaro Funakoshi Philipp Spanger Mikio Nakano Takenobu Tokunaga

[*1] （株）ホンダ・リサーチ・インスティチュート・ジャパン
Honda Research Institute Japan Co., Ltd.

[*2] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Department of Computer Science, Tokyo Institute of Technology

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. はじめに

参照表現の生成と理解は，音声対話システムなど自然言語を用いて人とコミュニケーションを行うシステムには欠かせない技術である．参照表現は，複数の対象の中から特定の対象を指示するために用いる情報によって二つに分類できる．すなわち，参照対象の属性を用いる対象記述（例えば「赤い机」）と，文脈と状況を用いる照応と直示（例えば「それ」）の二つである．本稿では，前者の対象記述について主に議論する．但し，本稿で提案する参照表現のモデルは，対象記述だけでなく照応と直示を扱うことも可能であり，その点が対象記述を扱うことを目的とした従来の手法と提案手法との差異の一つでもある．また，我々は生成と理解の両方に関与するモデルを提案するが，本稿では主に生成の視点から論述する．

過去２０年に渡り対象記述の生成に関する研究が行われてきた．この分野における基本的な課題は，複数の指示対象の中から特定の一つを唯一に参照する最小（使用する属性の数において）の記述を生成するアルゴリズムの開発であった．すなわち，対象記述生成の研究は最小性と唯一性の二つに制約されていたと言える．

しかしながら最小性の制約は緩和される方向にある（例えば，[Dale 95, Spanger 08]）．これは，厳密に最小性を追求すると計算量が増大すること，実際に人が生成する表現には冗長性がしばしば観察されること，またそのような表現の方がより自然であると人が感じることなどによる．これに対し，もう一つの制約である唯一性についてはあまり注意が払われてこなかった．本稿ではこの唯一性の制約を緩和することを提案する．

本研究の第一の目的は，物体を構成する様々な部位が異なる属性値を持つ不均一な物体に関する参照表現を扱うことである．参照表現生成に関する過去の研究では，図 1 中のテーブル A のような，均一な物体を扱うことが多かった．しかしながら，現実世界に存在するのはそのような単調な物体ばかりではない．テーブル B や C のような異なる特徴を持つ部分からなる不均一な物体は容易に見つかる．そこで本稿では「物体の部分」の概念を明示的に導入し，部分に関する言及を扱う．このとき，人が論理的には曖昧な表現を理解・生成することがあるため，唯一性の制約が問題となる．

例えば図 1 の状況でテーブル B を特定する目的で，人は「角の赤い机」のような表現を使う．論理的には，この表現はテーブル A と B の両方に等しく当てはまるため，唯一性の制約に違反している．しかしながら，人はこのような表現を頻繁に生成するし，それを聞いた人がテーブル A を選んだり，テーブル A と B の間で困惑するようなことはほとんどない（3. 節にて検証実験の結果を示す）．我々は，これは人の知覚特性に由来するある種の語用論的含意を反映していると考え，またそれは人の生成した表現を機械が理解したり，実環境において人が理解しやすい表現を機械が生成する上で重要であると考える．4. 節で提案するモデルは，確率に基づく定式化によって唯一性の制約を緩和し，人の知覚特性を考慮することで曖昧性を解決する．5. 節にて，提案モデルを用いた参照表現の理解と生成のアルゴリズムを示す．

現時点では，モデルに対する包括的評価を行うための十分なデータがない．そこで 6. 節では，研究の第一歩として，提案モデルが従来型のドメインモデリングを包含することを示すために，既存のコーパスを用いて行った評価実験の結果を示す．

最後に 7. 節で，提案モデルの利点や，重要だが本稿では扱わなかった様々な事項を，提案モデルによってどのように扱うことができるか考察する．

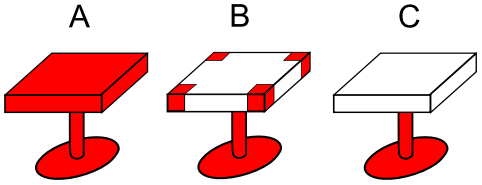

図 1: An example scene

連絡先: 船越 孝太郎，351–0188 埼玉県和光市本町 8-1 HRI-JP，
funakoshi@jp.honda-ri.com

2. 関連研究

[Horacek 05] では，エージェントの間での知識と認識の齟齬に起因する不確実性に対処するために確率を導入する事を提案している．我々のモデルは Horacek と問題意識を共有するが，我々の主眼は異なる所にある（すなわち，不均一な物体の取り扱いと唯一性の制約の緩和）．加えて，Horacek の提案手法は生成だけを扱っている．一方，我々のモデルは生成と理解の両方に関与する．[Roy 02] も参照表現生成のための確率モデルを提案しているが，均一な物体を前提としている．

[Horacek 06] では，文書のような構造を持った対象についての参照表現を扱っている．この研究では部分という概念を扱っているが，その研究の焦点は我々のそれとは大きく異なる．

3. 論理的唯一性に対する反例

図 1 に示した視覚刺激を用いて，参照表現の生成と理解に関する二つの心理学実験を行った．

一つ目の実験では，13 人の日本人被験者に「角の赤い机」という表現を提示し，その表現が図中のどのテーブルを指すかを訊ねた．結果，13 人中 12 人がテーブル B を選択した．また B を選択した 12 人の中の 7 人は，与えられた表現は曖昧ではないと答えた．

二つ目の実験では，最初の実験とは別の 13 人の日本人被験者にテーブル B を指示する表現を作るように指示した．但し，その際にテーブルの位置関係は用いないように指示した．13 人中 10 人は，一つ目の実験で被験者に提示した表現と同等の表現を回答した．13 人中 3 人だけが,「足と四つ角だけが赤い机」のように，論理的にもテーブル B を唯一に指し示す表現を回答した．

これらの結果は，状況によっては，人が論理的に曖昧な表現を容易に生成／理解することを示している．つまり，論理的な側面から考えれば曖昧であっても，人にとっては曖昧でない表現が存在するということである．

4. 提案モデル

ドメインに現れる k 種の「物体の部分」（部分のクラス）を $\pi = \{p^1, p^2, \ldots, p^k\}$ として定義する．ここで，p^1 は常にある物体の全体を意味する．家具のドメインでは，p^1 はその種類（椅子やテーブルなど）に関わらず，ある一つの家具を指す．その他の $p^i (i \neq 1)$ は,「脚」や「天板」のような家具の部分を表す．π はあるドメインに対して一つ定義されるものであり，一つ一つの物体に対して定義されるものではない．従って，物体によっては，p^i に相当する部分を持たないこともある（例えば，脚を持たない椅子もある）．

ある参照表現 e を属性値表現 e^a_j とそれが修飾する部分指示表現 e^p_j との n 個のペアで表す．

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

例えば,「赤い脚の白い机」は次のように表される．

$$\{(\text{机}, \text{白い}), (\text{脚}, \text{赤い})\}$$

物体の集合 ω と参照表現 e が与えられたとき，e が ω 中のある物体 o を指し示す主観確率を $\Pr(O = o|E = e, \Omega = \omega)$ で表す．もし現実により忠実なモデルを求めるのであれば，Ω に対しても確率分布を考えることができる．しかしここでは，Ω は ω に固定されている，すなわち関心のある物体の集合が話者間で正確に共有されていると考える．これにより，以降では $\Pr(o|e)$ は $\Pr(o|e, \omega)$ を意味するものとする．

定義 (1) に従い，確率 $\Pr(o|e)$ を次のように推定する．

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

ここで，$\mathcal{N}$ は正規化係数である．ベイズ則に従い，

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

と変形する．従って，

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

となる．$\Pr(e^p_i, e^a_i|o)$ は次のように展開する．

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

ここで，p_u は e^p_i によって指示される物体の部分であり，a_v は e^a_i によって表現される属性値を表す．おのおの属性値はなんらかの属性 α（属性値の集合）に属す．例えば，色という属性であれば $\alpha_{color} = \{red, white, \ldots\}$ のように定義される．また，$\Pr(a)$ は $\Pr(A^a = true)$ の省略表記である．つまり，属性値に関する確率変数は，属性毎ではなく属性値毎に二値変数として設定される．

ここで，部分指示表現 e^p_i と属性値表現 e^a_i はそれぞれ対応する物体の部分および属性値と一対一対応すると仮定すると，

$$\Pr(o|e) \approx \mathcal{N} \prod_i \frac{\Pr(o)\Pr(p_i, a_i|o)}{\Pr(p_i, a_i)} \tag{6}$$

$$\approx \mathcal{N} \prod_i \Pr(o|p_i, a_i) \tag{7}$$

となる．この一対一対応の仮定を置くことで，本稿では言語生成において語彙選択と呼ばれる問題を無視する．しかし提案モデル自体は，確率分布をコーパスから学習することで語彙選択を扱うことができる．

$\Pr(o|p, a)$ は参照表現生成における属性選択に関与する．従来研究で提案されている多くの属性選択アルゴリズムは，離散値をとる複数の属性に対する集合演算に基づく．この従来式のブール型ドメインモデリングに沿う最も単純な $\Pr(o|p, a)$ の推定は以下のように定式化できる．

$$\Pr(o|p, a) \approx \begin{cases} |\omega'|^{-1} & (o \text{ 中の } p \text{ が } a \text{ を持つ}) \\ 0 & (o \text{ 中の } p \text{ が } a \text{ を持たない}) \end{cases} \tag{8}$$

ここで ω' は，その要素である物体の部分 p が属性値 a を持つ，ω の部分集合である．

しかし，[Horacek 05] が指摘するように，多くの物理的性質は連続的であり，そのシンボル化は不確実性を生じるため，実環境においてはこの標準的なモデリングは十分ではない．例えば，二つの青い物体が有ったとして，一方は他方よりもより青いということがあり得る．ある人はある色を青というかもしれないが，他の人は紫だというかもしれない．これに加えて，1. 節で述べた論理的曖昧性の問題がある．すなわち，ある性質自体は複数の物体に同等に備わっていたとしても，それ以外の情

報（ここでは視覚的な文脈）が参照表現の解釈に影響する場合がある．

これらの現象は $\Pr(o|p,a)$ を次のように推定することで捉えられる可能性がある．

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

ここで，$\Pr(a|p,o)$ は物体 o の部分 p に関する属性値 a の関連性，すなわち当てはまりの程度を表す．例えば，a が赤色を意味する属性値で，$p = p^1$ だとすれば，物体全体が赤いテーブル A（図 1）に対しては 1 に近い確率値が，半分ほどしか赤くないテーブル C については 0.5 程度の値が与えられる．$\Pr(p|o)$ は物体 o 中における部分 p の顕現性，すなわち際立ちの程度を表す．これは「ある物体の部分が言及されたのであれば，そこは他の部分よりも目立っているはずである」というある種の語用論的含意を表現し，Grice の公理に関係する（参照表現と Grice の公理の関係については [Dale 95] に議論がある）．$\Pr(p|o)$ は，人間の視覚的注視のモデルとして提案されている顕現性マップ [Itti 98] などを用いて推定できる可能性がある．$\Pr(o)$ は o が選択される事前確率である．[Tokunaga 05] で用いられているようなポテンシャル関数を持ちいて $\Pr(o)$ を推定すれば，ある参照表現に等しく関連する複数の物体を話者との位置関係によって適切に順位付けることができるだろう．

1. 節で述べた論理的曖昧性の問題は，$\Pr(p|o)$ の導入によって多くの場合解決できると予想する．しかし，図 1 の例では，語用論的含意に関して別の解釈もできる．すなわち，『「あえてある部分の属性値が言及されたのであれば，そこ以外の部分はその属性値を持たない」という含意が存在し，それによって人はテーブル B を選択している』，という解釈もできる．このような含意を扱うには，$\Pr(o|p,a)$ を (9) とは違った形で推定するか，あるいは，情報検索における検索質問拡張のように，「脚の赤い白い机」という言語表現に対して $\{(\text{机}, \text{白い}), (\text{脚}, \text{赤い}), (\text{天板}, \text{赤くない})\}$ と含意を直接的に反映した内部表現を生成する方法が考えられる．

このように論理的曖昧性への対処には複数のアプローチが考えられるが，どれがよいのかを決めるには今後の検討を要する．

5. アルゴリズム

5.1 理解

参照表現 e の理解とは，物体集合 ω の中から指示対象 $\hat{o}$ を特定することである．これは，4. 節で提案したモデルを用いて次のように定式化できる．

$$\hat{o} = \underset{o \in \omega}{\operatorname{argmax}} \Pr(o|e). \tag{10}$$

5.2 生成

参照表現の生成とは，指示対象 $\hat{o}$ が与えられたときに，それを他の物体から区別する最も良い表現 $\hat{e}$ を選択することである．これは，次のよう定式化できる．

$$\hat{e} = \underset{e \in \rho}{\operatorname{argmax}} \Pr(e)\Pr(\hat{o}|e). \tag{11}$$

ここで，ρ は $\hat{o}$ について事前に生成された候補表現の集合である．候補集合の一番単純な生成方法は $\hat{o}$ について可能な全ての属性値を組み合わせる方法であるが，考慮する候補の数を減らすための工夫はいろいろと考えられる．

表 1: 生成結果（平均 Dice スコア）

	Furniture	*People*
提案手法	0.78	0.66
1st (IS-FBN)	0.80	0.74
5th (DIT-DS)	0.75	0.70

$\Pr(e)$ は物体とは独立した e の生成確率であり，分布はコーパスから推定できる．6. 節で説明する評価実験では，$\Pr(e)$ を次のように近似する．

$$\Pr(e) \approx \Pr(|e|)\prod_i \Pr(\alpha_i). \tag{12}$$

ここで，$\Pr(|e|)$ は言及された属性値の数として捉えた表現長の分布である．$\Pr(\alpha)$ は，特定の属性 α が選択される確率である（[Spanger 08] における $SP(a)$）．

6. 評価

先に述べたように，現時点では，モデルに対する包括的評価を行うための十分なデータが存在しない．そこで，本節では Generation Challenge 2009[*1] で Participant's Pack として提供されたデータを用いて，提案モデルが従来型のドメインモデリングを包含することを示す．使用したデータは European Language Resources Association (ELRA) を通じて一般公開されている TUNA コーパス[*2]の一部である．TUNA は，家具が配置されたドメイン（Furniture）と，人の上半身写真が配置されたドメイン（People）の二つのドメインにおいて，人が作成した対象記述表現を集めたコーパスである．Participant's Pack として配布されたデータは，training と development の二部からなる．training を確率分布の学習に用い，development を評価に使用した．

ここでは，$\Pr(o|p,a)$ に対し，均一な分布を仮定した．すなわち，(8) 式を用いて確率を計算した．まず，提案モデルを用いて参照表現理解（人が使用した属性値を元に指示対象を特定する）を実装した．Furniture ドメインと People ドメインの両方において，人がそもそも誤った表現を生成している場合をのぞくと，正解率は 100%であった．

次に，参照表現生成を実装し，人の生成した表現と我々の実装が生成した表現との類似度を Dice スコア [Belz 07] で測定した．Dice スコアは集合間の重なりの程度を 0 から 1 の間で表現する指標で，比較する二つの集合が同一の時に値が 1 となる．ここでは，参照表現に使用した属性値の集合を比較する．Furniture と People の両ドメインの結果を表 1 に示す．テストデータの数は Furniture で 80，People で 68 である．1st と 5th は 2007 年度の GRE-Challenge での上位 1 位と 5 位になった手法の結果を示している [Belz 07].

この評価実験における我々の主眼は提案モデルが従来の集合演算に基づくドメインモデリングを包含することを示すことであり，Dice スコアの値を最大化するためのパラメータ調整などには力を入れていない．しかし，結果として得られた Dice スコアは，2007 年度の GRE-Challenge に参加した 23 システムの中の上位 5 位から 7 位のシステム（[Belz 07] 参照）と同等の性能であることを示しており，比較的良い性能が得られたと言える．

∗1 http://www.nltg.brighton.ac.uk/research/genchal09/tuna/
∗2 http://catalog.elra.info/product_info.php?products_id=1074

7. 考察

従来研究では，物体の部分は無視されていたか，暗黙的に扱われていた．TUNA コーパスの場合であれば，Furniture ドメインでは物体の部分は無視されており，物体の弁別に関与しないように設計されていた．People ドメインでは，hair，glasses，beard など，部分の概念を含んでいたが，それらは hairColor のようにその部分の属性と結合させて，一つの属性として暗黙的に扱われていた．我々のモデルでは，物体の部分を明示的に扱うことで，部分の顕現性を表現し，論理的曖昧性の問題を扱うことができる．従来提案されているアルゴリズムでも何らかの拡張を行うことでこのような問題を扱うことができるかもしれないが，我々の定式化のほうがより明晰だろう．それに加え，我々のモデルは生成と理解の両方の基盤として用いることができる．

生成と理解は別々に研究されることが多いが，システムの統合を考えた場合，両者が一つの基盤の上に実装されていることが好ましい．一つにはそれによって実装と保守のコストを下げられるという理由がある．しかし，コスト以上の問題として，理解と生成を別々の基盤の上に実現すると，システムが理解できる表現と生成できる表現の間に容易に不一致が生じてしまう問題がある．このような対話システムはユーザにとって利用しづらく，良い印象を与えないであろう．理解できる表現と生成できる表現を一致させ，システムに一貫性を与えるには，両者が同一の基盤上にあることが望ましい．

1. 節の冒頭で述べたように，参照表現には物体の属性や関係を使用した記述表現と，指示詞や代名詞を用いた照応表現とがある．これらも従来は別々に研究されてきたが，我々のモデルに拡張を加えることで両者を統一的に扱うことができると考えている．記述表現の研究は対話文脈とは切り離してなされることがほとんどであるが，一旦対話の中に置かれれば記述表現と照応表現の間の境界は明確でなく連続的になる．従って，特に対話システムへの応用を考えた場合には，両者を統一的に扱う手法が重要になる．

本稿では物体間の「関係」（大小関係，位置関係など）を考慮しなかったが，我々のモデルを変更することなく，関係を扱う生成および理解のアルゴリズムを設計することはそれほど困難ではないと考える．また，我々のモデルを用いた参照表現生成は generate-and-test 方式になるため，計算量に懸念がある．しかし，我々のモデルでは唯一性の制約が緩和されているので，一定の計算時間内で唯一に指示する表現が必ずしも見つからなくても良い．従って，関連性や顕現性が高い限られた数の属性と部分だけを考慮することなどで，計算量を制御することは可能だろう．

従来，参照表現生成において個人差はほとんど考慮されてこなかったが，個人間での違いの大きさが改めて注目されてきている [Dale 09, Bohnet 09]．また，対話において相手が使用した表現に自分の表現を合わせる alignment という現象にも注目が集まっている [Buschmeier 09, Janarthanam 09]．提案モデルでは，どちらの現象に対しても，言語モデル $\Pr(e)$ や語彙選択確率 $\Pr(e_i^a|a_v, o)$ などを，話者と対話履歴に合わせて動的に調整することで対応できると期待できる．

今後の課題として，適切なドメインのデータを集め，語彙選択も含めたモデル全体の評価を行う必要がある．また，上記のような照応や関係の取り扱いについても取り組む予定である．

参考文献

[Belz 07] Belz, A. and Gatt, A.: 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)*, pp. 75–83 (2007)

[Bohnet 09] Bohnet, B.: Generation of Referring Expression with an Individual Imprint, in *Proc. the 12th European Workshop on Natural Language Generation (ENLG)*, pp. 185–186 (2009)

[Buschmeier 09] Buschmeier, H., Bergmann, K., and Kopptitle, S.: An Alignment-Capable Microplanner for Natural Language Generation, in *Proc. the 12th European Workshop on Natural Language Generation (ENLG)*, pp. 82–89 (2009)

[Dale 95] Dale, R. and Reiter, E.: Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions, *Cognitive Science*, Vol. 18, pp. 233–263 (1995)

[Dale 09] Dale, R. and Viethen, J.: Referring Expression Generation through Attribute-Based Heuristics, in *Proc. the 12th European Workshop on Natural Language Generation (ENLG)*, pp. 59–65 (2009)

[Horacek 05] Horacek, H.: Generating referential descriptions under conditions of uncertainty, in *Proc. the 10th European Workshop on Natural Language Generation (ENLG)* (2005)

[Horacek 06] Horacek, H.: Generating references to parts of recursively structured objects, in *Proc. the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (2006)

[Itti 98] Itti, L., Koch, C., and Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, pp. 1254–1259 (1998)

[Janarthanam 09] Janarthanam, S. and Lemon, O.: Learning Lexical Alignment Policies for Generating Referring Expressions for Spoken Dialogue Systems, in *Proc. the 12th European Workshop on Natural Language Generation (ENLG)*, pp. 74–81 (2009)

[Roy 02] Roy, D.: Learning Visually-Grounded Words and Syntax for a Scene Description Task, *Computer Speech and Language*, Vol. 16, No. 3 (2002)

[Spanger 08] Spanger, P., Kurosawa, T., and Tokunaga, T.: On "redundancy" in selecting attributes for generating referring expressions, in *Proc. the 22nd International Conference on Computational Linguistics (COLING)* (2008)

[Tokunaga 05] Tokunaga, T., Koyama, T., and Saito, S.: 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*, pp. 93 – 100 (2005)