

Bayesian Network による下位範疇化の確率モデルおよびその学習

宮田 高志, 宇津呂 武仁, 松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{takashi, utsuro, matsu}@is.aist-nara.ac.jp

格の依存関係と格要素の汎化レベルを考慮した下位範疇化の確率モデルを Bayesian Network として定式化し、局所的に最適なモデルを統計的に求めるアルゴリズムを実装・評価した。格および格要素の汎化レベルの間の依存関係を (1) 格同士の依存関係 (2) 汎化レベル同士の依存関係 (3) 格と汎化レベルの依存関係の三種類に分類し、格と汎化レベルの依存関係を同時に扱うモデルよりも格同士の依存関係を扱うモデルと汎化レベル同士の依存関係を扱うモデルを組み合わせたモデルの方が、曖昧性解消のタスクにおいては性能が良いことを実験的に確かめた。

[キーワード] 下位範疇化, 確率モデル, ベイジアンネットワーク, モデル学習

Probabilistic Models of Subcategorization Based on Bayesian Network and Their Learning from Corpus

MIYATA Takashi, UTSURO Takehito, MATSUMOTO Yuji
Graduate School of Information Science, Nara Institute of Science and Technology

We formalize two probabilistic models of verbal subcategorization based on the Bayesian network which treat dependencies among both cases and the class generalization of adjunct/argument nouns. We implement algorithms for obtaining locally optimal models and evaluate the resulting models in terms of syntactic disambiguation task. We categorize these dependencies into the three sorts; dependencies (1) among cases, (2) among class generalization, and (3) among both cases and class generalization. It is observed that the combined model of the model which treats dependencies among cases and the model which treats those among class generalization is superior to the model which treat dependencies among both cases and class generalization simultaneously.

[Keyword] subcategorization, probabilistic model, Bayesian network, model learning

1 はじめに

近年大規模コーパスの利用が盛んになるに従って、従来人手で作成していた様々な情報を自動もしくは半自動的に獲得・抽出するための研究が行われている。その中でも動詞の下位範疇化フレームは、名詞の分類や統語解析などに応用できるため、コーパスから自動的に獲得することは有用である。

ある動詞の下位範疇化フレームを特徴づける要素としては、格の種類と格要素の名詞クラスの二つがある。一般にこれらは互いに依存関係 (dependency) を持つ。すなわち、次の三種類の依存関係を考慮しなければならない。

- 格の種類同士の依存関係
(例) 同一種類の格は共起しにくい。

「*太郎 が 花子 が 読む」

- 格の種類と名詞クラスの依存関係
(例) 動作動詞の「が格」は生物をとりやすい。
「太郎 が 窓 を 開ける」¹
- 名詞クラス同士の依存関係
(例) 「と格」の名詞クラスは、同時に出現した格の名詞クラスと“近い”。
「太郎 と 花子 が 走る」

一方、複雑な依存関係を効率的に記述する枠組みとして Bayesian Network [9] が、主に意思決定や故障診断などの分野で利用されている。コーパスの大きさを変えたり他の言語現象との相互作用を考慮した

¹一つの格に関してだけではなく、二つ以上の格にまたがった依存関係もあり得る。(例) 同一種類の格であっても名詞クラスが異なれば共起することがある。「三時 に 東京 に 向かう」

りするときに定性的な議論を行えるようにするためには、このような確率的な枠組みに基づいたモデル化が重要である。本研究では上の三種類の依存関係を Bayesian Network を使ってモデル化し、MDL 原理 [11] に基づいてコーパスから自動的に依存関係を推定する方法を提案する。さらに曖昧性解消のタスクによって、いくつかの動詞に対して得られたモデルの評価を行う。

動詞の下位範疇化フレームに関する先行研究としてはまず、[10] や [7] がある。どちらも一つの動詞の一つの格について統語的曖昧性解消に最適な名詞クラスを探す研究で、依存関係に関しては考慮されていない。[8] では Dendroid 分布という、制限された Bayesian Network に基づいて依存関係をモデル化し、格の種類同士および格の種類と名詞クラスの依存関係について考察しているが、格の種類同士の間にしか統計的に有意な結果は得られなかったと報告している。ただし、格の種類と名詞クラスの依存関係を同時に扱っているのではなく、まず [7] で提案された方法で各格の名詞クラスを固定してからそれらの間の依存関係を検出するという方法をとっている。[4] は情報理論におけるデータ圧縮の考え方をを用いてコーパスからの下位範疇化フレームの学習を研究しているが、応用面での評価は行っていない。[14] では確率モデルではないが、いくつかのヒューリスティクスに基づいて格の種類と名詞クラスの依存関係を統計的に求めており、統語的曖昧性解消においてタスクの 74% から 92% に対して 72% から 97% の正解率を実現している。²

Bayesian Network の学習アルゴリズムに関する研究はすでに多数あり [2, 3, 1, 12, 6, 5, 13]、特に Dendroid 分布に関する学習アルゴリズムでは最適なモデルを見つけるためのアルゴリズムが [13] によって提案されている。一般に Bayesian Network は事象を頂点、依存関係を有向辺で表した directed acyclic graph (DAG) で表現されるが、[13] のアルゴリズムでは Dendroid 分布を前提としたことで仮定された、事象の間にあらかじめ決められた半順序を利用し、分枝制限法によって探索の効率化を図っている。事象の「意味」が明確で、あらかじめ依存関係に半順序を設けて制限しておくことができる場合には Dendroid 分布でも問題はないが、本研究のように「格が出現したかどうか」「格要素の名詞クラスは何か」を事象とする場合、このような仮定は妥当ではない。

以下ではまず、Bayesian Network に関する一般的な説明および学習アルゴリズムのガイドラインとし

て用いている MDL 原理の説明を行い、次に動詞下位範疇化の確率モデルについて説明する。最後に実験についての説明とその結果及び評価について述べる。

2 Bayesian Network

2.1 記法

$X = (x_1, x_2, \dots, x_N)$ を N 次元属性ベクトルの確率変数とする。添字集合 $I = \{1, 2, \dots, N\}$ の巾集合の要素 $\sigma = \{\sigma(1), \sigma(2), \dots, \sigma(k)\} \in 2^I$ に対して、 σX によって k 次元部分属性ベクトル $\sigma X = (x_{\sigma(1)}, x_{\sigma(2)}, \dots, x_{\sigma(k)})$ を定義する。例えば、 $\sigma = \{2, 4\}$ および $X = (x_1, x_2, x_3, x_4, x_5)$ の時、 $\sigma X = (x_2, x_4)$ である。

Bayesian Network は X の各要素 x_i を頂点とする DAG で表される。以下ではこのグラフを B_s と書く。 B_s が頂点 x_i から x_j への有向辺を持つ時、 x_i を x_j の親 (parent)、 x_j を x_i の子供 (child) という。ある頂点 x_i の親の集合を $\pi(x_i)$ と書くことにする。また頂点の集合 v から v 中の頂点の添字への関数を $\iota(v)$ と書くことにする。例えば $v = \{x_1, x_3, x_8\}$ ならば $\iota(v) = \{1, 3, 8\}$ である。頂点 x_i 及びその親 x_j に対応する属性がとり得る値の種類をそれぞれ k_i, k_j とする時、頂点 x_i には $(k_i - 1) \prod_{x_j \in \pi(x_i)} k_j$ 個のパラメータが割り振られる。各パラメータの意味は「親頂点に対応する属性の組合わせの条件の下で x_i に対応する属性がある値をとる、条件付き確率 (conditional probability)」である。 π および ι を使うと、頂点 x_i が持つパラメータ (条件付き確率) の集合は次のように書ける。

$$\bigcup_A \{ \Pr(x_i = a_i | \iota(\pi(x_i))X = \iota(\pi(x_i))A) \} \quad (1)$$

ここで A はすべての属性 $(x_1, x_2, \dots, x_N)$ のとり得る値のすべての組み合わせ $(a_1, a_2, \dots, a_N)$ にわたる。以下では各頂点に対するパラメータの割り当て方を B_p 、特に x_i に対するパラメータの集合を $B_p(x_i)$ と書く。また、グラフ構造 B_s および割り当て B_p を持つ Bayesian Network を $B = (B_s, B_p)$ と書く。図 1 に「雨が降っている」「傘をさした人がいる」「地面が濡れている」「気温が高い」の四つの属性を頂点として持つ Bayesian Network の例を示す。各属性はそれぞれ「雨が降っているかどうか (+/-)」「傘をさした人がいるかどうか (+/-)」「気温が高いか低いか (+/-)」の二通りおよび「地面が濡れているか普通か乾いているか (w/n/d)」の三通りの値をとり得るとする。例えば、 x_3 のパラメータのうち $\Pr(x_3 = w | x_1 = +)$ は「雨が降っている ($x_1 = +$) 時に地面が濡れている ($x_3 = w$) 確率」を表している。 $\Pr(x_3 = d | x_1 = +)$ が指定されていないのは $\sum_{a \in \{w, n, d\}} \Pr(x_3 = a | x_1 = +) = 1$ という制約があるからである。 $\Pr(x_1 = -)$ や $\Pr(x_2 = - | \dots)$ が指定されていないのも同様の理由による。

²[19] では Maximum Entropy に基づいて確率モデルとして定式化しなおしている。また、[16] では名詞クラスは考慮されていないが、構文解析器に組み込んだ時の精度について研究している。Maximum Entropy に基づいたモデルでは、多数の feature とそのパラメータによって依存関係を記述する。Feature は Bayesian Network の属性を複数組み合わせたものに相当するが、対応するパラメータにはその feature の重みという意味しかない (どの feature を用いるかを慎重に決めないと) 学習結果が直感的に把握しづらいという問題がある。

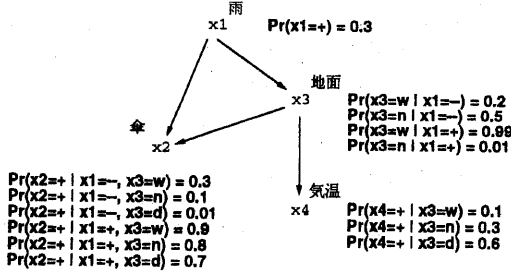


図 1: Bayesian Network の例

2.2 結合確率の計算

属性ベクトル $X = (x_1, x_2, \dots, x_n)$ がある値 $A = (a_1, a_2, \dots, a_n)$ をとる結合確率 (joint probability) を与えられた Bayesian Network B に基づいて計算するには、Bayes の定理を用いて次の値を求めればよい。

$$\Pr^{(B)}(X = A) = \prod_{x_i \in V} \Pr^{(B)}(x_i = a_i | \iota(\pi(x_i))X = \iota(\pi(x_i))A) \quad (2)$$

ここで V は B の頂点の集合である。例えば図 1 の Bayesian Network に基づいて「雨が降っておらず、傘をさした人がいて、地面が普通の状態、気温が高い」確率すなわち $\Pr((x_1, x_2, x_3, x_4) = (-, n, +, +))$ を計算すると、次のようになる。

$$\begin{aligned} \Pr((x_1, x_2, x_3, x_4) = (-, +, n, +)) &= \Pr(x_1 = -) \Pr(x_3 = n | x_1 = -) \times \\ &\quad \Pr(x_2 = + | x_1 = -, x_3 = n) \times \\ &\quad \Pr(x_4 = + | x_1 = -, x_3 = n, x_2 = +) \end{aligned} \quad (3)$$

$$\begin{aligned} &= \Pr(x_1 = -) \Pr(x_3 = n | x_1 = -) \times \\ &\quad \Pr(x_2 = + | x_1 = -, x_3 = n) \times \\ &\quad \Pr(x_4 = + | x_3 = n) \end{aligned} \quad (4)$$

$$= (1 - 0.3) \times 0.5 \times 0.1 \times 0.3 = 0.0105 \quad (5)$$

ここで $\Pr(x_4 = + | x_1 = -, x_3 = n, x_2 = +) = \Pr(x_4 = + | x_3 = n)$ であることに注意。これは、この Bayesian Network において x_4 は x_3 以外の属性には依存しないと定義されているからである。 B_p で指定されていない条件付き確率については複数の結合確率を使って間接的に計算される。例えば「気温が低いという条件の下での雨が降っている確率」すなわち $\Pr(x_1 = + | x_4 = -)$ は、次のように計算される。³

$$\Pr(x_1 = + | x_4 = -)$$

³属性同士の依存関係が複雑になってくると、この方法は計算量の点から現実的ではなくなってくる。大規模な Bayesian Network では、条件部の頂点に何らかの入力を与えて活性拡散を使って求める方法が用いられるが、格の間の依存関係はかなり簡潔なものであることが予想されたので、本研究では式 (6) のような方法をそのまま用いている。

$$\begin{aligned} &= \frac{\sum_{(a_2, a_3)} \Pr(x_1 = +, x_2 = a_2, x_3 = a_3, x_4 = -)}{\sum_{(b_2, b_3, b_4)} \Pr(x_1 = +, x_2 = b_2, x_3 = b_3, x_4 = b_4)} \quad (6) \\ &= \frac{0.3 \times (0.99 \times 0.9 + 0.01 \times 0.7 + 0)}{0.3} = 0.898 \quad (7) \end{aligned}$$

ただし、分子の (a_2, a_3) は $\{+, -\} \times \{w, n, d\}$ のすべての要素に、分母の (b_2, b_3, b_4) は $\{+, -\} \times \{w, n, d\} \times \{+, -\}$ のすべての要素にわたる。

2.3 MDL 原理と Bayesian Network における記述長

MDL 原理とは、確率モデルの選択における基準の一つで「与えられたデータにできるだけ忠実で、できるだけ簡潔なモデルを選ぶ」ように設計された基準であり、MDL 原理に基づくモデル選択とは、記述長 (description length) とよばれる量を最小にするようなモデルを選択することである。一般に与えられたデータ D に関する確率モデル M の記述長 $l(M, D)$ は、 M による D に対する対数尤度 $\log L_M(D)$ と、 M 中のパラメータの個数 N_M およびデータのサイズ $|D|$ の関数で、次のように定義される。[17]

$$l(M, D) = -\log L_M(D) + \frac{1}{2} N_M \log |D| \quad (8)$$

$l(M, D)$ を最小化するモデルとは、 $L_M(D)$ をできるだけ大きく (i.e. 与えられたデータにできるだけ忠実)、 N_M をできるだけ小さく (i.e. できるだけ簡潔) するようなモデルである。

N 個の頂点 $V = \{x_1, x_2, \dots, x_N\}$ からなる Bayesian Network $B = (B_s, B_p)$ の、データ集合 $D = \{A_1, A_2, \dots, A_M\}$ に対する記述長 $l(B, D)$ は次のように定義できる。

$$\begin{aligned} l(B, D) &= -\log_K \Pr^{(B)}(D) \\ &\quad + \frac{1}{2} \left(\sum_{x_i \in V} \prod_{x_j \in \pi(x_i)} |\text{dom}(x_j)| \right) \log_K M \end{aligned} \quad (9)$$

K は符号アルファベットの種類の数で、ここでは定数である。 $\text{dom}(x_j)$ は属性 x_j がとり得る値の集合を表し、右辺第二項中の $\sum_{i=1}^N \prod_{x_j \in \pi(x_i)} |\text{dom}(x_j)|$ はモデルに含まれているパラメータの個数を表す。右辺第一項はデータ集合 D の対数尤度であり、次のように計算される。

$$\begin{aligned} \log_K \Pr^{(B)}(D) &= \sum_{A \in D} \log_K \Pr^{(B)}(X = A) \\ &= \sum_{A \in D} \sum_{x_i \in V} \log_K P_i \end{aligned} \quad (10)$$

ただし、

$$P_i = \Pr^{(B)}(x_i = a_i | \iota(\pi(x_i))X = \iota(\pi(x_i))A)$$

3 下位範疇化の確率モデル

動詞の下位範疇化フレームを特徴づける属性として、格の種類およびシソーラス上の格要素のクラスを考えることにする。本研究では統語的な格を対象にしているので格の種類は助詞や前置詞で区別することにし、自明な依存関係⁴を極力避けるために用いるシソーラスは木構造を仮定する。以下では日本語を対象とするが、統語的に格の種類が区別できる言語であれば同様の確率モデルを構築することができる。

3.1 動詞・名詞共起の用例及び下位範疇化フレーム

動詞 v と共起した名詞の用例 $e^{(v)}$ を、次のような素性構造で表す。

$$e^{(v)} = \begin{bmatrix} p_1 : c_1 \\ p_2 : c_2 \\ \vdots \\ p_k : c_k \end{bmatrix} \quad (11)$$

ここで p_i は助詞を、 c_i はシソーラス上の葉頂点をそれぞれ表す。⁵例えば、「子供が空地で花を摘んだ」に対する素性構造は次のようになる。

$$e^{(\text{摘む})} = \begin{bmatrix} \text{が: 子供} \\ \text{で: 空地} \\ \text{を: 花} \end{bmatrix} \quad (12)$$

また、シソーラス上のクラス c_1 がクラス c_2 の下位クラスであるとき、 $c_1 \sqsubseteq c_2$ と書くことにする。

動詞 v に関する下位範疇化フレーム (subcategorization frame) $s^{(v)}$ とは、 v と共起する格に対する制約を記述した素性構造であり、次のように記述される。

$$s^{(v)} = \begin{bmatrix} p_1 : \begin{bmatrix} \text{pol: pol}_1 \\ \text{cnt: } c_1 \end{bmatrix} \\ p_2 : \begin{bmatrix} \text{pol: pol}_2 \\ \text{cnt: } c_2 \end{bmatrix} \\ \vdots \\ p_l : \begin{bmatrix} \text{pol: pol}_l \\ \text{cnt: } c_l \end{bmatrix} \end{bmatrix} \quad (13)$$

ここで pol_i は助詞 p_i が現れるかどうかを表す極性 (polarity) で + もしくは - の値をとる。意味制約 (sense restriction) c_i はその格要素がシソーラス上のどのクラスに属するかを表す。負の極性を持つ格 $\begin{bmatrix} p_i : \begin{bmatrix} \text{pol: -} \\ \text{cnt: } c_i \end{bmatrix} \end{bmatrix}$ は「助詞 p_i を持つ名詞クラ

ス c_i は、動詞 v と共起しない」ことを表す。また、二つの下位範疇化フレーム s_1 と s_2 を単一化したものを $s_1 \wedge s_2$ と書くことにする。

3.2 包摂関係

ある動詞 v の名詞との共起の用例 $e^{(v)}$ と v の下位範疇化フレーム $s^{(v)}$ が次の条件を満たすとき、 $e^{(v)}$ は $s^{(v)}$ に包摂 (subsume) されるといい、 $e^{(v)} \sqsubseteq s^{(v)}$ と書く。

1. $s^{(v)}$ 中の、- 極性を持たない全ての助詞 p_i に対して、
 - (a) p_i が + 極性を持つならば、同じ助詞 p_i が $e^{(v)}$ 中に存在する。
 - (b) p_i が名詞クラス c_i^s を持つてかつ同じ助詞 p_i が $e^{(v)}$ 中に存在するならば、対応する名詞 c_i^e が $c_i^s \sqsubseteq c_i^e$ を満たす。⁶
2. $s^{(v)}$ 中の、- 極性を持つ全ての助詞 p_i に対して、
 - (a) $e^{(v)}$ 中には同じ助詞 p_j が存在しないか、または
 - (b) p_i が名詞クラス c_i^s を持つてかつ $e^{(v)}$ 中の対応する名詞 c_i^e は $c_i^s \sqsubseteq c_i^e$ を満たさない。

以下では簡単のため、ある一つの動詞に着目した場合 v を省略することにする。

3.3 下位範疇化の優先度

今仮にある動詞に関して助詞と名詞クラスの組合わせが N 通りあるとすると、その動詞と共起する可能な格のパターンは 2^N 種類にもなるが、実際に観測されるのはそれらのうちの非常に少数である。そこで大多数の観測されない用例を制限するために依存関係を導入することが考えられる。例えば、「動詞「買う」の「が格」の名詞がクラス human で、「で格」の名詞がクラス money の時、「を格」の名詞はクラス object をとりやすい」は依存関係の一つを表現している。このような依存関係は、用例 e といくつかの (部分的な) 下位範疇化フレームとの包摂関係として (14) のように記述することができる。

$$e \sqsubseteq \begin{bmatrix} \text{が: } \begin{bmatrix} \text{pol: +} \\ \text{cnt: human} \end{bmatrix} \\ \text{で: } \begin{bmatrix} \text{pol: +} \\ \text{cnt: money} \end{bmatrix} \end{bmatrix} \Rightarrow e \sqsubseteq \begin{bmatrix} \text{を: } \begin{bmatrix} \text{pol: +} \\ \text{cnt: object} \end{bmatrix} \end{bmatrix} \quad (14)$$

用いる下位範疇化フレームの構造によって次の三通りの依存関係が考えられる。

⁴例えば二つの名詞クラス C_1 と C_2 に属する名詞の集合の間に $C_1 \subset C_2$ などの包含関係があった場合、内容に関係なく「 C_1 に属する」という属性は「 C_2 に属する」という属性に依存することになってしまう。

⁵名詞の意味に曖昧性がある場合については 5 節で述べる。

⁶ $e^{(v)}$ 中に同じ助詞が存在しなければ包摂することになる。名詞クラスだけが指定されている時は、用例中にその格が存在することを意味しないことに注意。

- 助詞 p_i の値として極性だけを許す。
(例) 「[が格]」と「[で格]」が出現したときは「を格」も出現しやすい」

$$e \subseteq \left[\begin{array}{l} \text{が: } [\text{pol: +}] \\ \text{で: } [\text{pol: +}] \end{array} \right] \Rightarrow e \subseteq [\text{を: } [\text{pol: +}]] \quad (15)$$

- 助詞 p_i の値として名詞クラスだけを許す。
(例) 「([が格]「[で格]」「[を格]」の三つの格が同時に出現したという条件の下で)「[が格]」の名詞クラスが human で、「[で格]」の名詞クラスが money の時は、「[を格]」の名詞クラスは object と取りやすい」

$$e \subseteq \left[\begin{array}{l} \text{が: } [\text{cnt: human}] \\ \text{で: } [\text{cnt: money}] \end{array} \right] \Rightarrow e \subseteq [\text{を: } [\text{cnt: object}]] \quad (16)$$

- 助詞 p_i の値として極性・名詞クラスの任意の組み合わせを許す。
(例は (14))

(15), (16), (14) を「 $e \subseteq s_1 \Rightarrow e \subseteq s_2$ 」というルールではなく、 $\Pr(e \subseteq s_2 | e \subseteq s_1) \Pr^w(e; s_1, s_2)$ という条件付き確率と単語の生成確率の積を用いれば、これらに応じて次の三種類の確率モデルを考えることができる。

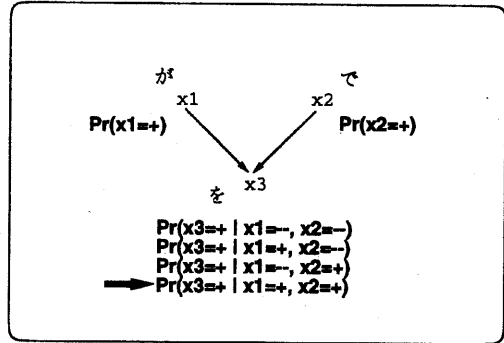
- 助詞 p_i を属性とし、各属性のとり得る値を + または - とするモデル (Slot Model)
- 助詞 p_i を属性とし、各属性のとり得る値を互いに排反な名詞クラス (i.e. シソーラス上の cut) とするモデル (Class Model)
- 助詞 p_i と名詞クラス c_i の対を属性とし、各属性のとり得る値を + または - とするモデル (Mixed Model)

ここで、単語の生成確率 $\Pr^w(e; s_1, s_2)$ とは下位範疇化フレーム s_1, s_2 で指定されている名詞クラスの中で、用例 e 中の単語が生成される確率を表す。本研究では e 中の単語 w_i が s_1 及び s_2 中の名詞クラス c_i で制限されているとき、 $L(c_i)$ をシソーラス上で c_i を上位クラスに持つ葉頂点の数として、 $\Pr^w(e; s_1, s_2)$ を $(\prod_i L(c_i))^{-1}$ で近似する。⁷ 図 2 及び図 3 に各モデルの例を示す。矢印がつけられた条件付き確率が (15), (16), (14) のルールに対応する。用例 e と下位範疇化フレーム s の間に包摂関係 $e \subseteq s$ が成り立っているかどうか、 s 中の素性値の指定すなわち属性の指定に帰着されていることに注意。

下位範疇化の優先度 (subcategorization preference) は、与えられた依存関係の下で用例 e が観測される確率で決められる。例えば図 3 の Mixed

⁷すなわち個別の単語に依らず、単語ごとに独立であると仮定する。

Slot Model



Class Model

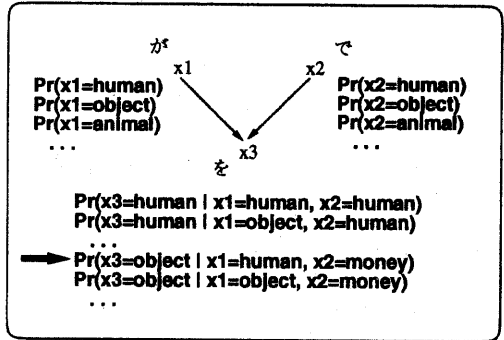


図 2: Slot Model と Class Model の例

Model では (明示的には) 六つの依存関係⁸が指定されているが、このような依存関係が成立している時に、ある用例 e が観測される確率を求めるには 2.2 節で述べた結合確率と単語の生成確率の積を計算すればよい。例えば図 3 の Mixed Model において、用例 $e = \left[\begin{array}{l} \text{で: お金} \\ \text{を: 花瓶} \end{array} \right]$ の優先度は $\frac{\Pr(x_2=+, x_3=+)}{L(\text{money})L(\text{object})}$ で計算される。

4 モデル推定

Slot Model, Class Model および Mixed Model はそれぞれ次の greedy search に基づいて推定できる。このアルゴリズムは、ネットワークの構造を DAG に保ちかつ記述長を減らすような変形を行うことでより良いモデルを構築してゆく。必ずしも最適なモデルが得られることは保証されないが、現在の所 exhaustive search は計算量の問題から実行不可能であるためこのアルゴリズムを用いた。⁹

⁸条件付きでない確率も特殊な依存関係(その格が条件なしで出現する)を表していることに注意。

⁹[13] での分枝限定法は効率化のひとつの方法である。なお、実装の都合上 5 節の実験と評価では Class Model の学習は行っていない。

Mixed Model

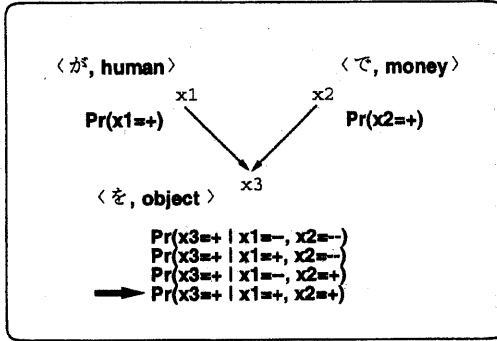


図 3: Mixed Model の例

1. $B_s^{(0)}$ を辺を一本も持たないグラフとし、コーパス $C = \{e_1, e_2, \dots, e_M\}$ に基づいて各 $B_p^{(0)}(x_i)$ を初期値に設定する。
2. $B_s^{(n)}$ に可能な変形を行ったグラフを $B_{s1}^{(n+1)}$, $B_{s2}^{(n+1)}$, ..., $B_{sk}^{(n+1)}$ とする。
3. 各 $B_{si}^{(n+1)}$ に対して、
 - (a) コーパス C に基づいて各頂点のパラメータ $B_{pi}^{(n+1)}(x_j)$ を求める。
 - (b) 記述長 $l(B_{si}^{(n+1)}, C)$ を求める。ただし、Class Model と Mixed Model に関しては 3.3 節で説明したように、単語の生成確率を考慮する必要がある。すなわち式 (9) 及び式 (10) の定義で、 P_i を次のように変更した記述長を用いる。

$$\frac{Pr^{(B)}(x_i = a_i | \iota(\pi(x_i))X = \iota(\pi(x_i))A)}{L(c_i)} \quad (17)$$

4. $l(B_{si}^{(n+1)}, C)$ がもっとも小さいグラフを一つ選び、それを改めて $B_{si}^{(n+1)}$ とおく。
5. $l(B_{si}^{(n+1)}, C)$ が $l(B_{si}^{(n)}, C)$ より小さくなければ終了。

各モデルにおける最初のグラフ $B_s^{(0)}$ の頂点、 $B_s^{(n)}$ から $B_{si}^{(n+1)}$ への変形¹⁰および変形によって新しく作られたモデルに関するパラメータの割り当て $B_{pi}^{(n+1)}$ は次のように定義される。ただし、簡単のため条件 $\iota(\pi(x_j))X = \iota(\pi(x_j))A$ は $X_j = A_j$ と略記し、 $C(x_i = a_i, x_j = a_j, \dots)$ はコーパス C において、条件 $x_i = a_i, x_j = a_j, \dots$ を満たす用例の出現回数を表すとする。

Slot Model: $B_s^{(0)}$ の頂点はコーパス C 中に現れた格の助詞とし、各頂点 (属性) のとり得る値は +

または - とする。変形は任意の二つの頂点の間に有向辺を張ることとし、 n ステップ目のパラメータの割り当てを表 1 式 (18) とする。¹¹

Class Model: $B_s^{(0)}$ の頂点はコーパス C 中に現れた格の助詞とし、各頂点 (属性) のとり得る値はシソーラス上の根とする。変形は任意の二つの頂点の間に有向辺を張るか、もしくはある頂点のとり得る値をより細かくする (i.e. $\text{dom}(x_j)$ のうちの一つの名詞クラス c_k をその下位クラスで置き換える)¹²こととし、 n ステップ目のパラメータの割り当てを表 1 式 (19) とする。¹³

Mixed Model: $B_s^{(0)}$ の頂点はコーパス C 中に現れた格の助詞 p_j とシソーラス上の根 r の対 (p_j, r) とし、各頂点 (属性) のとり得る値は + または - とする。変形は任意の二つの頂点の間に有向辺を張るか、もしくはある頂点 (p_j, c_j) を、名詞クラス c_j の下位クラス $c_{j1}, c_{j2}, \dots, c_{jk}$ に従って k 個に分割する¹⁴こととし、 n ステップ目のパラメータの割り当てを表 1 式 (20) とする。

5 実験・評価

実験には EDR Japanese bracketed corpus [18] を用いた。この中から動詞「買う」に関して 510 個の、「住む」に関して 507 個の名詞との共起の用例をそれぞれ抽出した。これらに含まれていた助詞の種類は約 30 種類であった。

シソーラスとしては約 45,000 語からなる分類語彙表 [15] を用いた。用例中の名詞クラスが一意に定まらないときはすべての可能性を列挙して元の用例の頻度をそれらの間で分配した。例えば、二つの名詞 X と Y がそれぞれ $\{C_1, C_2\}$ と $\{D_1, D_2, D_3\}$ の曖昧性を持つとき、用例 $\begin{bmatrix} p_1: X \\ p_2: Y \end{bmatrix}$ を $\begin{bmatrix} p_1: C_i \\ p_2: D_j \end{bmatrix}$ ($i \in \{1, 2\}, j \in \{1, 2, 3\}$) の六つに分け、それらの頻度を $1/6$ とする。

5.1 評価方法

得られた確率モデルを評価するために、格の係り先として二つの動詞が有り得るときにその曖昧性を解消するタスクを用いて、正解率を測定した。次の構造を持った文を考える。

$$N_s - p_s - N_{11} - p_{11} - \dots - N_{1m} - p_{1m} - V_1 \\ - N_{21} - p_{21} - \dots - N_{2n} - p_{2n} - V_2. \quad (21)$$

¹¹実際には親頂点の集合が変化した頂点に関してのみ、更新すればよい。他のモデルについても同様である。

¹²この場合はネットワークの構造は変化しない。

¹³厳密に言えば各パラメータの条件部分には「 x_j, X_j のすべての格が出現している」という条件がつく。

¹⁴この時、頂点 (p_j, c_j) は辺を持たないとする。頂点の分割に伴って辺の分割まで考えると可能な分割の仕方が膨大になるため、ここでは分割を行うのは辺を持たない頂点だけに限定する。また、分割する代わりにシソーラスの葉頂点からはじめて併合してゆくことも考えられる。

¹⁰どの変形も、結果が DAG である場合のみ適用可能とする。

$$B_p^{(n)}(x) = \left\{ \Pr(x_j = + | X_j = A_j) = \frac{C(x_j = +, X_j = A_j)}{C(X_j = A_j)} \mid A_j \in \{+, -\}^{|\pi(x_j)|} \right\} \quad (18)$$

$$B_p^{(n)}(x_i) = \left\{ \Pr(x_j = c_k | X_j = A_j) = \frac{C(x_j = c_k, X_j = A_j)}{C(X_j = A_j)} \mid c_k \in \text{dom}(x_j), A_j \in \prod_{x_i \in \pi(x_j)} \text{dom}(x_i) \right\} \quad (19)$$

$$B_p^{(n)}(x_i) = \left\{ \Pr(\langle p_j, c_j \rangle = + | X_j = A_j) = \frac{C(\langle p_j, c_j \rangle = +, X_j = A_j)}{C(X_j = A_j)} \mid A_j \in \{+, -\}^{|\pi(x_j)|} \right\} \quad (20)$$

表 1: 各モデルにおける新しいパラメータの割り当て

ここで N_{π} と N_{ij} は名詞、 p_{ij} は助詞、 V_1 と V_2 は動詞であり、最初の動詞 V_1 は従属節をなす。格 $N_{\pi} - p_{\pi}$ が V_1 と V_2 のどちらに係るかを決定するのが、この実験のタスクである。例えば次の文では「自転車で」が $N_{\pi} - p_{\pi}$ に相当する。

(太郎が) 自転車で 西へ 走る 花子 を
駅 まで 追い掛けた。

問題の作成方法は以下の通りである。まず、各動詞に関する二つの用例 e_1^+ と e_2^- を取り出す(図4左)。次に e_1^+ から任意の格 $p_{\pi} : N_{\pi}$ をランダムに一つ取り出し、 e_2^- へ移動させる(図4右)。用例 e_1^+ の優先度を $\Pr(e_1^+)$ とする時、正しい対 (e_1^+, e_2^-) に対する優先度を $\Pr(e_1^+) \Pr(e_2^-)$ 、移動後の間違った対 (e_1^-, e_2^+) に対する優先度を $\Pr(e_1^-) \Pr(e_2^+)$ で定義する。正しい対の優先度の方が大きい、すなわち $\Pr(e_1^+) \Pr(e_2^-) > \Pr(e_1^-) \Pr(e_2^+)$ ならば正解とする。

優先度を計算するための確率モデルとしては Slot Model と Class Model の組合せおよび Mixed Model の二つについて調べた。

Slot+Class Model 下位範疇化フレームを、極性に関するフレーム s_s と名詞クラスに関するフレーム s_c に分けて、優先度を計算する。

$$\begin{aligned} & \Pr(e \sqsubseteq s_s \wedge s_c \mid e \sqsubseteq s'_s \wedge s'_c) \\ &= \Pr(e \sqsubseteq s_s \mid e \sqsubseteq s'_s \wedge s'_c) \times \\ & \quad \Pr(e \sqsubseteq s_c \mid e \sqsubseteq s'_s \wedge s'_c, e \sqsubseteq s_s) \\ &= \Pr(e \sqsubseteq s_s \mid e \sqsubseteq s'_s) \times \\ & \quad \Pr(e \sqsubseteq s_c \mid e \sqsubseteq s'_c, e \sqsubseteq s_s \wedge s'_s) \end{aligned} \quad (22)$$

具体的には一つのコーパスから学習した二種類の Bayesian Network を用いて式 (22) の右辺の各項を別々に計算し、その積を優先度とする。ただし今回は実装の都合上 Class Model の学習は行わず、「買う」に関しては「が格」と「を格」、「住む」に関しては「が格」と「に格」の間の依存関係だけを考慮した。この時の各格の cut は人手で求めた。

Mixed Model 4 節のアルゴリズムと式 (14) を用いて学習した Bayesian Network に基づいて優先度を計算する。

正解率/ 適用度 (%)	Slot	Slot+Class	Mixed
Basic	52.84/66.30	67.82/69.63	44.75/66.95
Heuristic ($T' = 0.05$)	67.20/66.30	78.55/69.63	47.22/69.28
Heuristic ($T' = 0.01$)	74.04/67.43	80.31/68.51	49.28/68.91

表 2: 「買う」及び「住む」に対する実験結果

また使い方についても、確率モデルだけを用いる場合と、ヒューリスティクスと組み合わせる場合の二通りについて調べた。

Basic 確率モデルに基づいて計算した優先度の積 $\Pr(e_1^+) \Pr(e_2^-)$ と $\Pr(e_1^-) \Pr(e_2^+)$ の差がある閾値 T を越えていれば、これらの大小に従って判断する。

Heuristic トレーニングに用いた用例を、得られたモデルの名詞クラスによって抽象化しておき、それらの頻度を表にしておく。テスト用の四つの用例 $e_1^+, e_1^-, e_2^+, e_2^-$ の頻度をそれぞれ $f_1^+, f_1^-, f_2^+, f_2^-$ とおく。問題の格が共起するかしないかでこれらの頻度に十分大きな差がある、すなわち $f_1^+ (1 - f_2^-)$ と $(1 - f_1^-) f_2^+$ がどちらもある閾値 T' を越えていれば、(確率モデルは使わずに) これらの大小に従って判断する。そうでなければ確率モデルによって判断する。

5.2 評価結果

Class Model に関しては人手で与えたために、移動させる格は「が格」「を格」「に格」の三種類だけに限定して実験を行った。表 2 に各々のモデルの適用度及び正解率を掲げる。¹⁵これらは「買う」の用例 510 個および「住む」の用例 507 個をそれぞれ 10 個のグループに分け、9 個のグループで (Slot 及び Mixed Model を) 学習し、残りの 1 個でテストをしたときの 10 回の平均である。優先度の閾値 T は、各モデルの適用度がほぼ同じ (66%~69%) になるようにモデルごとに調節した。理論的には等価な

¹⁵例えば Slot Model は Basic において、タスクの 66.3% 中 52.84% を正解したことを表す。

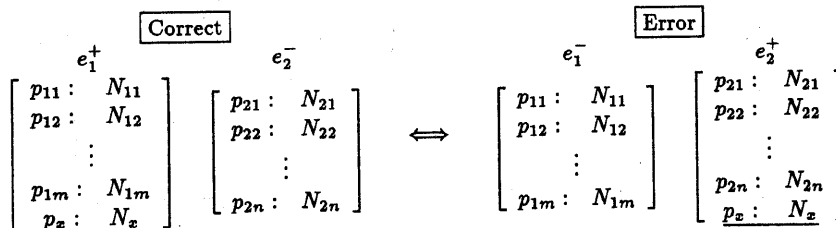


図 4: テスト用共起データの生成方法

Slot+Class Model と Mixed Model に差が現れたのは、次のような理由からだと思われる。まず、多くの格は単独にしか出現しない。そのために格の出現に関しては「この格が現れるときはこの格は現れにくい」という負の依存関係が強く現れることになる。よって名詞クラスを一緒に扱うと、名詞クラスの依存関係が相対的に小さくなってしまい、適切なモデルが得られないということになる。

6 結論

本研究では格の依存関係と格要素の汎化レベルを考慮した、動詞の下位範疇化のための確率モデルを Bayesian Network として定式化した。その際、依存関係を三種類に分類し、それぞれに関して greedy search に基づいたモデル推定アルゴリズムを実装し、二つの動詞「買う」と「住む」の曖昧性解消タスクによって評価した。その結果、格および格要素の汎化レベルの間の依存関係を同時に扱うモデルよりも分離して扱うモデルの方が、曖昧性解消のタスクにおいては性能が良いことを実験的に確かめた。

参考文献

- [1] Remco R. Bouckaert. "Probabilistic Network Construction Using the Minimum Description Length Principle". In Michael Clarke, Rudolf Kruse, and Serafin Moral, editors, *Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Vol. 747 of *LNCS*, pp. 41-48. Springer-Verlag, 1993. (Proceedings of European Conference ECSQARU'93, Granada, Spain, November 8-10, 1993).
- [2] C. K. Chow and C. N. Liu. "Approximating Discrete Probability Distributions with Dependency Trees". *IEEE Transactions on Information Theory*, Vol. 14, No. 4, pp. 462-467, 1968.
- [3] Gregory F. Cooper and Edward Herskovits. "A Bayesian Method for the Induction of Probabilistic Networks from Data". *Machine Learning*, Vol. 9, pp. 309-347, 1992.
- [4] M. Haruno. "Verbal Case Frame Acquisition as Data Compression". In *Proceedings of the Fifth International Workshop on Natural Language Understanding and Logic Programming (NLULP5)*, pp. 45-50, 1995.
- [5] D. Heckerman, D. Geiger, and D. M. Chickering. "Learning Bayesian Networks: The Combination of Knowledge and Statistical Data". *Machine Learning*, Vol. 20, No. 2, pp. 197-243, 1995.
- [6] W. Lam and F. Bacchus. "Learning Bayesian Belief Networks: An Approach Based on the MDL Principle". *Computational Intelligence*, Vol. 10, No. 3, pp. 269-293, 1994.
- [7] Hang Li and Naoki Abe. "Generalizing Case Frames Using a Thesaurus and the MDL Principle". In *Proceedings of International Conference on Recent Advances in Natural Language Processing*, pp. 239-248, Bulgaria, September 1995.
- [8] Hang Li and Naoki Abe. "Learning Dependencies between Case Frame Slots". In *Proceedings of the 16th International Conference on Computational Linguistics (COLING '96)*, Vol. 1, pp. 10-15, Copenhagen, Denmark, August 1996.
- [9] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, Los Altos, CA, 1988.
- [10] P. Resnik. "Semantic Classes and Syntactic Ambiguity". In *Proceedings of the Human Language Technology Workshop*, pp. 278-283, 1993.
- [11] Jorma Rissanen. "Universal Coding, Information, Prediction, and Estimation". *IEEE Transactions on Information Theory*, Vol. IT-30, No. 4, pp. 629-636, July 1984.
- [12] Joe Suzuki. "A Construction of Bayesian Networks from Databases Based on an MDL Principle". In *Proceedings of Uncertainty in AI*, pp. 266-273, 1993.
- [13] Joe Suzuki. "Learning Bayesian Belief Networks Based on the Minimum Description Length Principle: An Efficient Algorithm Using the B&B Technique". In *Proceedings of the International Conference on Machine Learning*, pp. 462-470, 1996.
- [14] Takehito Utsuro and Yuji Matsumoto. "Learning Probabilistic Subcategorization Preference by Identifying Case Dependencies and Optimal Noun Class Generalization Level". In *Proceedings of Applied Natural Language Processing*, pp. 364-371, Washington, DC, USA, April 1997.
- [15] 国立国語研究所. "分類語彙表", 1964, 1993.
- [16] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. "最大エントロピー法による格の従属関係の学習". 第三回年次大会発表論文集, pp. 337-340, 京都大学, March 1997. 言語処理学会.
- [17] 韓太舜, 小林斤吾. 情報と符号化の数理, 岩波講座 応用数学, 第 13 巻, 複合情報源のユニバーサル符号化, pp. 211-248. 岩波書店, 東京, December 1994.
- [18] 日本電子化辞書研究所. "EDR 電子化辞書仕様説明書", 1995.
- [19] 宇津呂武仁, 宮田高志, 松本裕治. "最大エントロピー法による下位範疇化の確率モデル学習および統計的曖昧性解消による評価". 情報処理学会研究会資料. 情報処理学会, May 1997. (to appear).