2P1-11in1

ベイジアンネットを用いた言語野モデルの構築に向けて

Toward the construction of the language area model using a Bayesian net

尾崎 竜史　一杉 裕志
Ryushi Ozaki　Yuuji Ichisugi

産業技術総合研究所 人工知能研究センター
AIST Artificial Intelligence Research Center

We believe that the cerebral cortex areas share a common principle of operation and the principle is some kind of Bayesian network. Based on the belief, we are aiming at a model of language area in the human brain based on a restricted and high-speed Bayesian network. As the first step towards that, we construct a system that can parse a very simple grammar based on a Bayesian network.

1. はじめに

大脳皮質が行う情報処理のモデルとして有望視されているものの一つにベイジアンネットがある [Lee and Mumford 03][Ichisugi 07]. 比較的複雑な離散確率分布をベイジアンネットで表現することができる．また，これにより柔軟な情報処理を行うことができる．われわれは，脳の言語処理機構をベイジアンネットとして模倣することを試みた．2. で範疇文法とその構文解析を説明し，3. では範疇文法の構文解析をベイジアンネットで行う方法について説明する．

2. 範疇文法とその構文解析

人間が言語を発したり理解したりする認知能力それ自体についての関心は二十世紀に入ってから現在に至るまで活発に研究されている．以下のような英文を考える：

(a) The sun shines.

(b) The sun whistles.

(c) Perhaps horse if will however shine.

容易にわかるように，文 (a) は完全に正しい英文であり，(b) は *sun* と *whistles* の結びつきが意味をなさないが構文的には正しい英文である．これに対し，(c) に現れる単語はすべて正当な英単語であるにもかかわらず文として不成立である，あるいは同じことだが，構文的に誤っている．言い換えれば，ヒト（読者や著者）には，３つの文 (a)(b)(c) を，順に「構文的に正しく，意味をなす文」「構文的に正しいが，意味をなさない文」「構文的に誤った文（非文）」と判断する認知能力がある．

2.1 範疇文法

構文的に正しい文を判別する機構について考察において，Ajdukiewicz は意味についての現象学的な観点に基づき，単語に基礎範疇あるいは関数範疇を割り当てた．そして，これらの範疇属性に基づいて単語同士が結びつくルールを設けることで自然言語の断片を構成した [Ajdukiewicz 1935]．彼の枠組みでは範疇 X を受け取って範疇 Y を返す関数範疇として $\frac{Y}{X}$ だけが扱われたが，この体系は後に拡張され，関数範疇として右から引数を取るもの Y/X と左から引数を取るもの $X\backslash Y$ が区別されて [*1] 扱われるようになった [Bar-Hillel 1953, Lambek 1958]. 関数範疇を作る操作を繰り返すことにより，基礎範疇の集合 $\mathcal{B}$ から，より複雑な範疇の集合 $\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, \ldots$ が次のように構成される：

$$\mathcal{C}_0 := \mathcal{B}, \tag{1}$$

$$\mathcal{C}_{i+1} := \mathcal{C}_i \cup \left\{ Y/X, X\backslash Y \;\middle|\; X, Y \in \mathcal{C}_i \right\} \quad (i \geq 0). \tag{2}$$

範疇の集合 $\mathcal{C}_n$ から $\mathcal{C}_{n+1}$ を作る操作は，理論的には何度でも繰り返すことができるが，以下では $\mathcal{C}_2$ までしか扱わない．表記を簡単にするため，以下では単に $\mathcal{C}$ と書いた場合は $\mathcal{C}_2$ を意味するものと約束しておく．

範疇 X, Y, Z から得られる複合的な関数範疇

$$(X\backslash Y)/Z$$

を，以下では

$$X\backslash Y/Z$$

のように略記することにする．

基礎範疇の集合 $\mathcal{B}$ としてどのようなものを考えることが適切であるかを一般的な考察のみから決定することは不可能である [Ajdukiewicz 1935]．以下では Ajdukiewicz に従い、基礎範疇の集合を

$$\{S, N\} \tag{3}$$

に取る．以下では，表 1. のように単語とその範疇を予め与えておく．

単語	範疇
cat, fish	N
black, white	N/N
eats	$N\backslash S/N$

表 1: 今回の文法で扱う単語とその範疇

連絡先: 尾崎 竜史，産業技術総合研究所 人工知能研究センター，ozaki.ryushi@aist.go.jp

*1 [Lambek 1958] の記号法に基づく．左から範疇 X を受け取って範疇 Y を返す関数範疇を表すために $Y\backslash X$ を使う流儀が採用されることもある．

2.2 解析木

範疇文法における構文解析の手順を説明するために，構文木を扱うための用語を定義しておく．単語の集合を W とする．$n(\geq 1)$ 個の終端ノードを持つ構文木の集合 $\mathcal{T}_n$ を式 (4)(5) によって再帰的に定義する：

$$\mathcal{T}_1 := W \times \mathcal{C}, \tag{4}$$

$$\mathcal{T}_n := \bigcup_{i=1}^{n-1} (\mathcal{T}_{n-i} \times \mathcal{T}_i) \times \mathcal{C} \qquad (n \geq 2). \tag{5}$$

終端ノードの個数を特定しない構文木の集合 $\mathcal{T}$ を次のように定義する：

$$\mathcal{T} := \bigcup_{n=1}^{\infty} \mathcal{T}_n.$$

以下の議論のために，次が成り立つことに念のため注意しておく：

$$W \times \mathcal{C} \ \subseteq \ \mathcal{T}.$$

構文木 $\langle w, T\rangle \in \mathcal{T}$ を以下ではしばしば

$$w{:}T$$

と表す．

2.3 範疇文法の構文解析

範疇文法の構文解析は，構文木の有限列

$$w_1{:}T_1,\ w_2{:}T_2, ..., w_n{:}T_n$$

に対する，次の二つの推論規則 (6)(7) として定義される：

$$(>A)\ \frac{\Gamma,\ w_1{:}T_2/T_1,\ w_2{:}T_1,\ \Delta}{\Gamma,\ \langle w_1{:}T_2/T_1, w_2{:}T_1\rangle{:}T_2,\ \Delta}, \tag{6}$$

$$(<A)\ \frac{\Gamma,\ w_1{:}T_1,\ w_2{:}T_1\backslash T_2,\ \Delta}{\Gamma,\ \langle w_1{:}T_1, w_2{:}T_1\backslash T_2\rangle{:}T_2,\ \Delta}. \tag{7}$$

水平線の上部が入力列であり，下部が出力列である．また，Γ, Δ は構文木からなる任意の有限列である．これらの操作を繰り返して適用し，最後に $w{:}S$ の形の構文木が一個だけ残ったら終了する（S は基礎カテゴリーの一つである (3)）．規則 (6)(7) の繰り返しによって得られた $w{:}S$ の形の構文木を，以下では「正格な構文木」と呼ぶ．また，正格な構文木の集合を $\mathcal{T}'$ で表す．

ここに現れる操作 (6)(7) は Γ, Δ の取り方に任意性がありうる．また，Γ, Δ を適切に取ることが不可能になって終了前に操作が続行できなくなる可能性がある．これらの理由により，以上の手続きは非決定的なものとなる．

3. ベイジアンネットによる構文解析器

CKY 法は，文脈自由言語の構文解析をボトムアップで行うために開発された手法である [Kasami 1965][Younger 1967][Cocke and Schwartz 1970]．簡単に言えば，これは三角行列の対角成分に入力単語列の文法記号を割り当て，与えられた文脈自由文法の生成規則にしたがって表を埋めてゆくことで構文木を構成する方法である．

N 個の単語からなる文を受理する CKY パーサを構成する三角行列の $N(N+1)/2$ 個の要素を，ベイジアンネットのノードに置き換えることにより，範疇文法のパーサを，範疇を値として持つベイジアンネットとして作成することができる．

ノード $A_{p,q}$ $(1 \leq p \leq q \leq N)$ に対して，親ノードの集合を次のように定めることによりベイジアンネットの構造を定義する：

$$\mathrm{par}\,(A_{p,q}) = \{\, A_{t,q} \mid 1 \leq t < p \,\} \cup \{\, A_{p,u} \mid q < u \leq N \,\}. \tag{8}$$

今回は 0 と非 0 だけを区別するような条件付き確率表を持つような簡易版のベイジアンネットを実装した．

構文木集合 $\mathcal{T}_N$ の範疇情報を $N(N-1)$ 個のノード $A_{i,j}$ $(1 \leq i \leq j \leq N)$ への割り当てを考える．集合 X の要素からなる $N \times N$ の上三角行列の集合を $\mathrm{Up}(X; N)$ とする．$*$ を範疇集合 $\mathcal{C}$ に属さない任意の元とする．関数

$$\Phi^N \colon \mathcal{T}_N \to \mathrm{Up}(\{*\} \cup \mathcal{C}; N) \ \ (N = 1, 2, ...) \tag{9}$$

を次のように式 (10)(11) によって再帰的に定義する：

$$\Phi^1(w{:}T)_{1,1} := T, \tag{10}$$

$N \geq 2$ のとき

$$\Phi^N(\langle t, u\rangle{:}T)_{i,j} := \begin{cases} T & i = 1, j = N \\ \Phi^M(t)_{i,j} & 1 \leq i \leq j \leq M \\ \Phi^{N-M}(u)_{i-M,j-M} & M < i \leq j \leq N \\ * & \text{otherwise} \end{cases}. \tag{11}$$

ただし，(11) において $t \in \mathcal{T}_M$, $u \in \mathcal{T}_{N-M}$ $(1 \leq M < N)$ とする．

このようにして定義された関数 Φ_N によって，構文木 $x \in \mathcal{T}_N$ に対して

$$\Pr\Big(A_{i,j} = \Phi^N(x)_{i,j} \mid A_{p,q} = \Phi^N(x)_{p,q}\ \ (A_{p,q} \in \mathrm{par}(A_{i,j}))\Big) := \begin{cases} \text{non-zero} & x \in \mathcal{T}', \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

として各ノードの条件付き確率表を定義する．

表 1. に示されるような語彙を用意し，この語彙から作られる $N = 4$ の正格な構文木の集合に対して上のような手順で $N = 4$ のベイジアンネットを作成し，このベイジアンネットが受理可能な文をすべて生成させたところ 16 個の文が得られ，いずれも構文的に正しい文であった．以上の結果から，CKY パーサの構造を参考にして構成したベイジアンネットによって単純な英文法の構文解析が可能であることが確認された．

作成されたベイジアンネットの構成図とこのベイジアンネットが受理する文の例を図 1 に示す．

4. まとめと今後の課題

4.1 まとめ

CKY パーサの構造を参考にして構成したベイジアンネットを構成し，これによって単純な英文法の構文解析が可能であることを確認した．

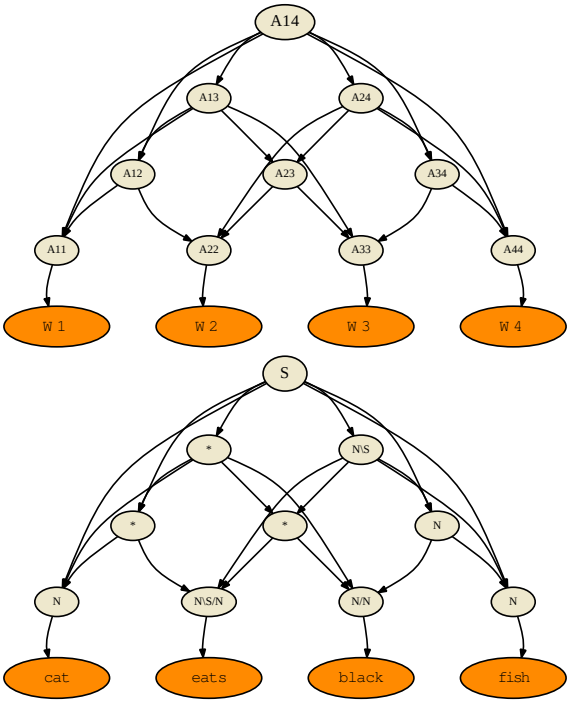

図 1: (上) 今回構成したベイジアンネットの構造．(下) 今回構成したベイジアンネットが受理した文の一つと，そのときの各ノードの範疇値．記号「∗」は有効な範疇が割り当てられていないことを示す．

4.2 今後の課題

今回作成したベイジアンネットは簡易実装のため，条件付き確率表に適切な重み付けを行っていない．今後は条件付き確率表を正確に実装し，確率文脈自由文法を実装することを目指す．

範疇文法はその後拡張され，ラムダ項による意味表現を扱うことができるようになっている [Steedman and Baldridge 2007][Bekki 2010]．このような方向への拡張も検討中である．

ベイジアンネットによる構文解析を扱ったものとして [Pynadath and Wellman 98] が知られている．今回のわれわれの実装を含む，素朴なベイジアンネットの実装においては，親ノードの数に対して計算量が指数関数的に増大するためスケーラビリティが低い．脳がベイジアンネットであるという仮説を信頼する場合，ヒトの脳が十分な速度で動くことには何らかの説明が必要である．一つの可能な説明は，脳を構成しているベイジアンネットが構造や演算の能力において何らかの制約が存在し，それにより解の探索空間のサイズが小さくなっていることである．noisy-OR 演算 [Pearl 88] はそのような制約の候補である．このようなベイジアンネットの計算法の変更により，より大きな入力に対する構文解析が高速に行えることを確かめることも今後の課題である．

謝辞

議論を通じて有用な示唆を与えてくださった佐藤泰介氏，戸次大介氏，高橋直人氏に感謝いたします．

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO） の委託業務の結果得られたものです．

参考文献

[Ajdukiewicz 1935] Ajdukiewicz, K: “Die syntaktische Konnexität”, Studia Philosophica, vol.1, pp.1-27 (1935); (English tlanslation: Storrs McCall(ed), “Polish Logic 1920-1939”, Oxford University Press, pp.203-231 (1967))

[Bar-Hillel 1953] Bar-Hillel, Y: “A Quasi-Arithmetical Notation for Syntactic Description”, Language, vol.29, no.1, pp.47-58 (1953)

[Lambek 1958] Lambek, J: “The Mathematics of Sentence Structure”, The American Mathematical Monthly, vol.65, no.3, pp.154-170 (1958)

[Lee and Mumford 03] Lee, T. S., and Mumdord, D.: “Hierarchical Bayesian inference in the visual cortex”, Journal of Optical Society of America A, vol.20, no.7, pp.1434-1448 (2003)

[Ichisugi 07] Ichisugi, Y.: “A Cerebral Cortex Model that Self-Organizes Conditional Probability Tables and Executes Belief Propagation”, International Joint Conference on Neural Networks, pp.178-183 (2007)

[Pynadath and Wellman 98] Pynadath, D. V. and Wellman, M. P.: “Generalized Queries on Probabilistic Context-Free Grammers”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.10, no.1, pp.65-77 (1998)

[Pearl 88] Pearl, J.: “Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference”, Morgan Kaufmann (1988)

[Kasami 1965] Kasami, T.: “An efficient recognition and syntax-analysis algorithm for context-free languages”, (Technical report). AFCRL. pp.65-758 (1965)

[Younger 1967] Younger, D. H.: “Recognition and parsing of context-free languages in time n^3”, Information and Control, vol.10, no.2, pp.189208 (1967)

[Cocke and Schwartz 1970] Cocke, J. and Schwartz, J. T.: “Programming languages and their compilers: Preliminary notes”, Technical report, Courant Institute of Mathematical Sciences, New York University (1970)

[Steedman and Baldridge 2007] Steedman, M. and Baldridge, J.: “Combinatory Categorial Grammar”, in : “Non-Transformational Syntax”, Blackwell (2007)

[Bekki 2010] Bekki, D.: “日本語文法の形式理論 — 活用体系・統語構造・意味合成 —”, くろしお出版 (2010)

[Pynadath 1996] Pynadath, D. V and Wellman, M. P. “Generalized queries on probabilistic context-free grammars”, Proceedings of the 14th National Conference on Artificial Intelligence (AAAI ’96), pp. 1285-1290 (1996)