大規模格フレームに基づく構文・格解析の統合的確率モデル

河原 大輔　　黒橋 禎夫
東京大学 大学院情報理工学系研究科
{kawahara, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

近年、構文解析は高い精度で行うことができるようになった。構文解析手法は、ルールベースのもの (e.g. [3])、統計ベースのもの (e.g. [5]) に大別することができるが、どちらの手法も基本的には、形態素の品詞・活用、読点や機能語の情報に基づいて高精度を実現している。例えば、

(1)　弁当を食べて出発した

という文は、「弁当を → 食べて」のように正しく解析できる。これは、「〜を」はほとんどの場合もっとも近い用言に係るという傾向を考慮しているからである。このような品詞や機能語などの情報に基づく係り受け制約・選好を、ルールベースの手法は人手で記述し、統計ベースの手法はタグ付きコーパスから学習している。しかし、どちらの手法も語彙的な選好に関してはほとんど扱うことができない。

(2) a.　弁当を出発する前に食べた
　　b.　弁当は食べて出発した

(2a) では、「弁当を」が (1) と同じように扱われ、「弁当を → 出発する」のように誤って解析される。(2b) においては、「〜は」が文末など遠くの文節に係りやすいという傾向に影響されて、やはり「弁当は → 出発した」のように誤って解析されてしまう。これらの場合、「弁当を 食べる」のような語彙的選好が学習されていれば正しく解析できると思われる。統計的構文解析器においては多くの場合、語彙情報が素性として考慮されているが、それらが用いている数万文程度の学習コーパスからでは、データスパースネスの影響を顕著に受け、語彙的選好をほとんど学習することができない。

さらに、2 項関係の語彙的選好が十分に学習されたとしても、次のような例を解析することは難しい。

(3)　太郎が食べた花子の弁当

「弁当を 食べる」「花子が 食べる」という語彙的選好を両方とも学習しているとすると、「食べた」の係り先はこれらの情報からでは決定することができない。この例文を正しく解析するには、「食べた」は「太郎が」というガ格をもっており、ヲ格の格要素は被連体修飾詞「弁当」であると認識する必要がある。このように、語彙的選好を述語・項構造としてきちんと考慮できれば構文解析のさらなる精度向上が期待できる。

述語・項構造を明らかにする格解析を実用的に行うためには、語と語の関係を記述した格フレームが不可欠であり、それもカバレージの大きいものが要求される。我々は、コーパスから大規模格フレームを自動的に構築する手法を提案してきた [1]。本稿では、この大規模格フレームに基づく構文・格解析の統合的確率モデルを提案する。本モデルは、格解析を生成的確率モデルで行い、格解析の確率値の高い構文構造を選択するということを行う。

2　構文・格解析の統合的確率モデル

本稿で提案する構文・格解析統合モデルは、入力文がとりうるすべての構文構造に対して確率的格解析を行い、もっとも確率値の高い格解析結果をもつ構文構造を出力する。すなわち、入力文 S が与えられたときの構文構造の確率 $P(T|S)$ を最大にするような構文構造 T_{best} を出力する。

$$T_{best} = \underset{T}{\operatorname{argmax}}\, P(T|S) \quad (1)$$

上式は、ベイズ則により 2 つの確率の積に分割できる。

$$T_{best} = \underset{T}{\operatorname{argmax}}\, P(T)P(S|T) \quad (2)$$

$P(T)$ は、構文構造 T を生成する確率である。$P(S|T)$ は、入力構造 T を仮定したときに各述語・項構造 PA_i を生成する確率の積とし、次のように定義する。

$$P(S|T) = \prod_{\mathrm{PA}_i \in T} P(\mathrm{PA}_i) \quad (3)$$

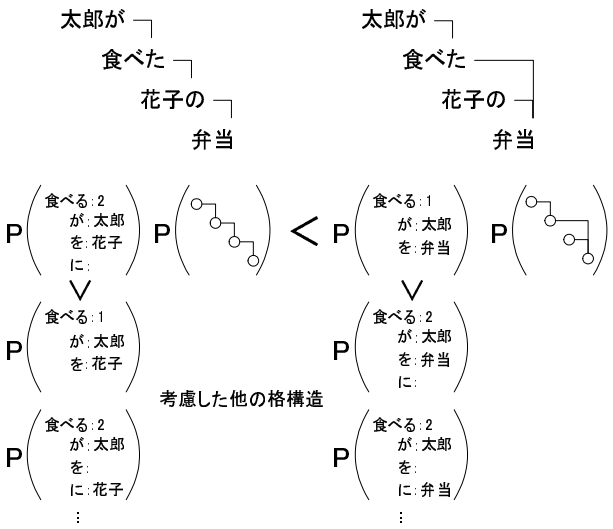

図 1: 構文・格解析統合モデルの概略図

図 1 に本手法の概略図を示す。この図は、文 (3) の解析時の例であり、右側の構文構造がもっとも確率値が高いために選択されることになる。そのときに使われた格フレーム「食べる:1」とその述語・項構造より、被連体修飾詞「弁当」が「食べた」に対してヲ格の関係をもつことも分かる。

以下では、述語・項構造の確率モデルと構文構造の確率モデルについて順に述べる。

2.1 述語・項構造の確率モデル

述語・項構造の生成モデルは、その述語・項構造にマッチする格フレームの選択と、入力側の各格要素の格フレームへの対応付けを同時に行うモデルである。

述語・項構造の生成確率 $P(\mathrm{PA}_i)$ は、述語 v_i を生成する確率、述語 v_i から格フレーム F_l を生成する確率、格フレーム F_l から格の対応関係 C_k を生成する確率の 3 つの積とする。

$$P(\mathrm{PA}_i) \quad = \quad P(v_i)P(F_l|v_i)P(C_k|F_l) \qquad (4)$$

格の対応関係 C_k とは、図 2 に示すように、入力側の格要素と格フレームの格との対応付け全体を表す。対応関係は図示のもの以外にも、「弁当」をニ格に対応付けるものなど様々な可能性がある。

$P(F_l|v_i)$ を格フレーム生成確率と呼び、その推定については 2.1.1 節で述べる。以下では、$P(C_k|F_l)$ について詳説する。

格の対応関係 C_k を、格フレームの格スロット s_j ごとに考える。格スロット s_j に入力側の格要素 (語 w_j, 表層格 c_j*) が対応付けられているかどうかで場合分けすると、次のように書き換えることができる。

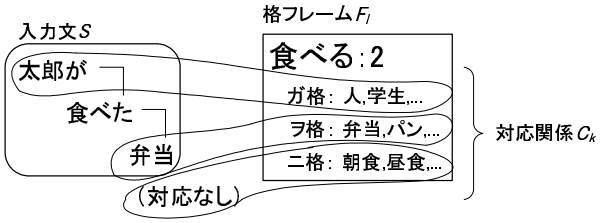

図 2: 格の対応関係 C_k の例

$$P(C_k|F_l) = \prod_{s_j:A(s_j)=1} P(A(s_j)=1, w_j, c_j|F_l) \cdot \prod_{s_j:A(s_j)=0} P(A(s_j)=0|F_l) \qquad (5)$$

ただし、$A(s_j)$ は、格スロット s_j に入力側格要素が対応付けられていれば 1、そうでなければ 0 をとる関数である。

右辺第 1 項の各確率は次のように分解できる。

$$P(A(s_j)=1, w_j, c_j|F_l) = P(A(s_j)=1|F_l) \cdot P(w_j, c_j|F_l, A(s_j)=1) \qquad (6)$$

この $P(A(s_j)=1|F_l)$ と式 (5) の $P(A(s_j)=0|F_l)$ を格生成確率、$P(w_j, c_j|F_l, A(s_j)=1)$ を用例生成確率と呼ぶ。これらの推定については、2.1.2 節と 2.1.3 節で述べる。

例えば、次の文を考える。

(4) 弁当は食べる

「食べる」のある格フレーム F_l がガ格とヲ格をもっているならば、この格フレームを用いたときの述語・項構造生成確率としては、「弁当は」をガ格またはヲ格に対応付けるときの 2 つを考えることになる。

$$\begin{aligned} P_{l1}(\text{弁当は食べる}) &= P(\text{食べる}) \cdot P(F_l|\text{ 食べる}) \\ &\cdot \quad P(\text{弁当}, \text{は}, A(\text{を})=1|F_l) \\ &\cdot \quad P(A(\text{が})=0|F_l) \\ P_{l2}(\text{弁当は食べる}) &= P(\text{食べる}) \cdot P(F_l|\text{ 食べる}) \\ &\cdot \quad P(\text{弁当}, \text{は}, A(\text{が})=1|F_l) \\ &\cdot \quad P(A(\text{を})=0|F_l) \end{aligned}$$

このような確率を「食べる」のすべての格フレームについて考え、もっとも確率の高い格フレームとそのときの格の対応関係に決定する。

*表層格とは、格要素末尾にある「が」「を」「に」「は」「では」「には」などの助詞列を表す。

述語・項構造の確率モデルは、式 (4) の $P(\mathrm{PA}_i)$ を最大にする格フレーム F_l と格の対応関係 C_k を求めるというモデルであるが、$P(v_i)$ は定数であるので、実際には次の確率を考えることにする。

$$P'(\mathrm{PA}_i) = P(F_l|v_i)P(C_k|F_l) \quad (7)$$

2.1.1 格フレーム生成確率の推定

格フレーム生成確率 $P(F_l|v_i)$ を推定するために、シソーラスによる類似度計算に基づく格解析 [1] を大規模コーパスに適用し、その格解析結果を用いる。用言 v_i の頻度と、v_i が格フレーム F_l をとる頻度を計数し、次のように最尤推定を行う。

$$P(F_l|v_i) = \frac{C(F_l, v_i)}{C(v_i)} \quad (8)$$

2.1.2 格生成確率の推定

前節と同じ格解析結果から、格フレームの使用頻度と格フレームの各格の出現頻度を計数する。格生成確率は、それらの頻度を用いて最尤推定を行う。

$$P(A(s_j) = 1|F_l) = \frac{C(A(s_j) = 1, F_l)}{C(F_l)} \quad (9)$$

$$P(A(s_j) = 0|F_l) = 1 - P(A(s_j) = 1|F_l) \quad (10)$$

2.1.3 用例生成確率の推定

格要素の語 w_j と表層格 c_j を生成する確率は独立であり、表層格の解釈は格フレームに依存しないと考えられるので、用例生成確率を以下のように近似する。

$$P(w_j, c_j|F_l, A(s_j) = 1) \approx P(c_j|s_j) \cdot P(w_j|F_l, A(s_j) = 1) \quad (11)$$

$P(c_j|s_j)$ は、表層格を解釈した格をタグ付けした関係コーパス [2] を用いて、次のように推定する。

$$P(c_j|s_j) = \frac{C(c_j, s_j)}{C(s_j)} \quad (12)$$

$P(w_j|F_l, A(s_j) = 1)$ は、自動構築した格フレームにおける用例の頻度を用いて、次のように推定する。

$$P(w_j|F_l, A(s_j) = 1) = \frac{C(w_j, F_l, s_j)}{C(F_l, s_j)} \quad (13)$$

2.2 構文構造の確率モデル

構文構造の確率モデルは、次のような日本語文の構文構造の特徴を考慮する。

- 近い文節に係る傾向がある
- 提題助詞をもつ文節は、文の主節や「〜が」「〜ので」などの強い区切れとなる節に係る傾向がある

これらの特徴に基づき、構文構造確率 $P(T)$ は、係り受け関係をもっている各文節ペア $(b_i \rightarrow b_j)$ の距離 d_{ij} と提題属性 t_{ij} を考慮する。文節 b_i, b_j のもつ素性を f_i, f_j とし、$P(T)$ を次のように定義する。

$$P(T) = \prod_{b_i \rightarrow b_j} P(d_{ij}, t_{ij}|f_i, f_j) \approx \prod_{b_i \rightarrow b_j} P(d_{ij}|f_i) \cdot P(t_{ij}|f_i) \quad (14)$$

距離 d_{ij} は、b_j が b_i の係り先候補文節の何番目にあたるかとし、d_{ij} に影響を与える f_i としては、読点の有無 p_i、表層格 c_i、提題助詞の有無 t_i の 3 つを考慮する。従って、$P(d_{ij}|f_i)$ は次のように表せる。

$$P(d_{ij}|f_i) = P(d_{ij}|p_i, c_i, t_i) \quad (15)$$

提題属性 t_{ij} は、b_i が提題助詞をもっている場合における、係り先 b_j が属する節の区切れとしての強さとする。節の強さとしては、南による節の分類 [7] を参考にして設定した 5 段階を考える。f_i としては、読点の有無 p_i と提題助詞の有無 t_i を考慮する。従って、$P(t_{ij}|f_i)$ は次のように表せる。

$$P(t_{ij}|f_i) = P(t_{ij}|p_i, t_i) \quad (16)$$

ただし、b_i が提題助詞をもたない場合 $(t_i = 0)$ は一定、つまり $P(t_{ij}|p_i, t_i = 0) = 1/5$ とする。

$P(d_{ij}|p_i, c_i, t_i)$ と $P(t_{ij}|p_i, t_i = 1)$ は、正解の構文構造がタグ付けされている京大コーパス [4] から推定する。

$$P(d_{ij}|p_i, c_i, t_i) = \frac{C(d_{ij}, p_i, c_i, t_i)}{C(p_i, c_i, t_i)} \quad (17)$$

$$P(t_{ij}|p_i, t_i = 1) = \frac{C(t_{ij}, p_i, t_i = 1)}{C(p_i, t_i = 1)} \quad (18)$$

3 実験

提案手法によって解析した構文構造の評価実験を行った。格フレームは新聞記事 26 年分 (約 2,600 万文) から自動構築し、用例生成確率はそれから推定した。格フレーム生成確率と格生成確率は新聞記事 26 年分の格解析結果から推定した。実験は、京大コーパス中の 1,653 文[†]を形態素解析器 JUMAN に通した結果

[†]これらの文は、格フレーム構築と各モデル学習には用いていない。

表 1: 構文解析の精度

	全体	連体修飾節
KNP	10870/12040 (90.3%)	1367/1456 (93.9%)
旧手法	10821/12040 (89.9%)	1369/1456 (94.0%)
本手法	10878/12040 (90.3%)	1381/1456 (94.8%)

を提案システムに入力し、文節の係り受け評価を行った。文末から 2 つ目の文節以外の係り受けすべてと連体修飾節の評価結果を表 1 に示す。表において、KNP とは、格解析を伴わない構文解析器 KNP による精度であり、旧手法は、シソーラスを利用した格解析の評価値に基づいて構文構造を選択する手法 [1] による精度である。

マクネマー検定を行った結果、本手法の精度は旧手法より有意 ($p = 0.0097 < 0.01$) に上回っていることがわかったが、KNP に対しては有意差はなかった。以下に、旧手法では誤りになるが、本手法によって正解になった例を挙げる。四角形で囲まれた文節の係り先が×下線部から○下線部に変化したことを示している。

- 奇跡的な経済発展を もたらした 官僚主導$_{\times}$ のシステム$_{\bigcirc}$ が、逆に障害となり機能不全に陥っているのです。
- 神戸市が 専門知識を 持つ$_{\times}$ 民間ボランティアを 募集した$_{\bigcirc}$ ところ、各地から大変な人数の応募があった。

4 関連研究

これまでに、語彙的選好を明示的に扱う構文解析手法がいくつか提案されてきた。白井らは、PGLR の枠組みに基づく統計的構文解析手法を提案している [8]。語彙的選好として、例えば $P(\text{パイ} \mid \text{を}, \text{食べる})$ のような確率を新聞記事 5 年分から学習しているが、用言の意味的曖昧性は扱っていない。京大コーパス中の比較的短い 500 文を用いて評価を行い、84.34%の解析精度であったと報告している。

藤尾らは、語の共起確率に基づく構文解析手法を提案している [6]。2 つの語が係り受けをもつ確率と距離確率の積で定義した確率モデルを用いており、それらの確率は EDR コーパスから学習している。EDR コーパス 1 万文を用いて評価を行い、86.89%であったと報告している‡。

‡文末から 2 つ目の文節も評価に入れている。

一方、語彙情報を素性として用いている様々な機械学習手法が提案されている。もっとも良い精度を実現しているのは工藤らの手法であり、彼らは係り先候補間の係りやすさを相対的に考慮する統計的構文解析手法を提案している [5]。最大エントロピー法を用いてモデル化し、京大コーパスから学習している。同コーパス中の 9,278 文を用いて評価を行い、91.37%の精度を実現している‡。しかし、数万文程度のタグ付きコーパスからでは、係り先候補間の語彙的選好を十分学習するのはほとんど困難であると思われる。

5 おわりに

本稿では、大規模格フレームに基づく構文・格解析の統合的確率モデルを提案した。このモデルによって、構文解析の精度が向上することを確認した。今後は、省略解析を統合することによって、格フレームに基づく構文・格・省略解析の統合的確率モデルを構築する予定である。

参考文献

[1] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431, 2002.

[2] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013, 2002.

[3] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, pp. 507–534, 1994.

[4] Sadao Kurohashi and Makoto Nagao. Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources and Evaluation*, pp. 719–724, 1998.

[5] 工藤拓, 松本裕治. 相対的な係りやすさを考慮した日本語係り受け解析. 情報処理学会 自然言語処理研究会 2004-NL-162, pp. 205–212, 2004.

[6] 藤尾正和, 松本裕治. 語の共起確率に基づく係り受け解析とその評価. 情報処理学会論文誌, Vol. 40, No. 12, pp. 4201–4212, 1999.

[7] 南不二男. 現代日本語文法の輪郭. 大修館書店, 1993.

[8] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. 統計的構文解析における構文的統計情報と語彙的統計情報の統合について. 自然言語処理, Vol. 5, No. 3, pp. 85–106, 1998.