大規模語彙的知識に基づく
構文・並列・格構造解析の統合的確率モデル

河原 大輔　　黒橋 禎夫
情報通信研究機構　京都大学大学院情報学研究科
dk@nict.go.jp　　kuro@i.kyoto-u.ac.jp

1　はじめに

並列構造の曖昧性は、構文的曖昧性のひとつであり、文章の意味理解に大きく影響を与えるため、並列構造の解析は非常に重要である。並列構造解析は、次の 2 つのタスクからなる。

- 並列構造の検出
- 並列構造の範囲の同定

並列構造の検出は、並列構造を導く表現が文章中に出現しても、それが常に並列構造を導くわけではないため必要である。例えば「〜と」は並列名詞句を導く表現であるが、用言のト格としても使われるという曖昧性がある。次に、並列構造の範囲、つまりどの句・節が並列になっているのかを同定する必要がある。

並列構造解析は、基本的に並列構造の類似性を用いる手法が提案されてきた [1, 6]。黒橋らは、並列構造の類似性に基づき、検出と範囲同定の両方を行う手法を提案している [6]。この手法は、並列構造候補の類似度が閾値を越えた場合に並列と判定し、もっとも類似度が高くなる部分を並列構造と同定するものである。しかし、類似性のみを用いているために、次のような例を解析できないという問題がある。

(1) a. 彼女とお金をやりとりした
　　b. 彼女とお金を大事にした

「彼女」と「お金」は、(1a) では並列ではなく、(1b) では並列である。どちらの例も「彼女」と「お金」の部分が共通であり、類似度に差が生じないため、類似性のみではこれらを区別することができない。このような曖昧性を解消するためには、用言の格フレームが必要となる。つまり、「やりとりする」がト格をとり、「大事にする」がト格をとらないという知識、また「大事にする」のヲ格が「彼女」や「お金」をとるという知識があれば、(1b) のみ「彼女」と「お金」が並列になっていることがわかる。

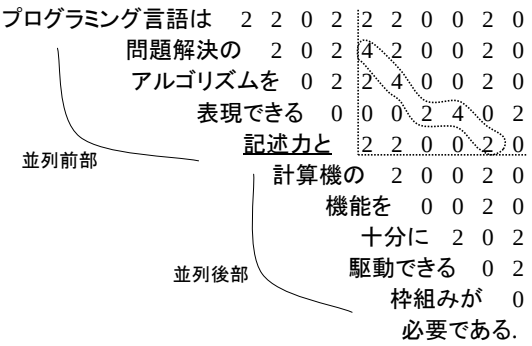


図 1: 類似度計算の例

本論文では、並列構造解析を格フレームに基づく確率的構文・格構造解析に統合する手法を提案する。提案手法は、並列構造候補の類似性、大規模コーパスから獲得した格フレームおよび並列関係の共起情報を利用し、構文・並列・格構造を同時に求める確率モデルとなっている。

2　並列句・節候補間の類似度計算

本論文では、並列構造の存在を示す表現を並列キーと呼び、並列構造の前半 (並列構造の開始点から並列キーまで) を並列前部、後半を並列後部と呼ぶ。並列キーは、読点や「と」「や」「および」などの付属語をもつ文節である。

並列構造解析においては、並列前部と並列後部の表現の類似性が手がかりとなる。類似性は、並列句・節候補の主辞単語間だけではなく、並列構造全体に現れるので、並列構造全体の類似性を求める黒橋らの手法 [6] を採用した。以下では、この手法について説明する。

類似度計算の例を図 1 に示す。まず、任意の文節間の類似度を、品詞のマッチ、単語のマッチ、シソーラス上での類似度に基づいて求める。シソーラスとしては、分類語彙表 増補改訂版 [5] を用いた。この類似度は、図 1 では行列の各要素に表現されている。次に、文節間の類似度の和に基づいて、任意の文節列間の類似度スコアを計算する。1 文節が複数の文節に対応する場合にはペナルティを課す。また、スコアを文節列全体の長さによって正規化する。黒橋らの手法は、図 1 の点線部分のように、閾値を越え、もっともスコアが高い文節列を並列構造に決定するものである。本研究では、あらゆる並列構造候補に対する類似度スコアの計算のみを行い、次節で説明する統合モデルの中で利用する。

3 構文・並列・格構造解析の統合的確率モデル

本研究では、依存構造に基づく確率的生成モデルを提案する。本モデルは、入力文 S が与えられたときの構文構造 T と格構造 L の同時確率 $P(T, L|S)$ を最大にするような構文構造 T_{best} と格構造 L_{best} を出力する。次のように、$P(S)$ は一定であるので、本モデルは $P(T, L, S)$ を最大にすることを考える。

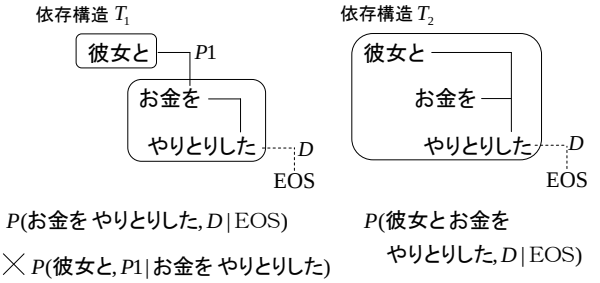
$$\begin{aligned}(T_{best}, L_{best}) &= \underset{(T,L)}{\operatorname{argmax}} P(T, L|S) \\ &= \underset{(T,L)}{\operatorname{argmax}} \frac{P(T, L, S)}{P(S)} \\ &= \underset{(T,L)}{\operatorname{argmax}} P(T, L, S)\end{aligned}$$

本モデルは節を基本単位とし、主節 (文末の節) から順次生成していく。ここで節とは、述語を 1 つ含みそれに関係する格要素群を含む部分 (述語項構造)、連体修飾句、並列前部の句の 3 種類と考える。$P(T, L, S)$ は、節 C_i を生成する確率の積として次のように定義する。

$$P(T, L, S) = \prod_{i=1..n} P(C_i, rel_{ih_i}|C_{h_i})$$

n は文 S 中に存在する節の数であり、C_{h_i} は節 C_i の係り先の節である。主節 C_n は係り先をもたないが、係り先を $C_{h_n} = \text{EOS}$ とする。rel_{ih_i} は C_i と C_{h_i} の係り受け関係を表し、通常の係り受け (D) または並列 (P)、並列の場合には並列類似度により細分化され、以下の 6 値をとるものとする。

$$rel_{ih_i} = \{D, P0, P1, P2, P3, P4\}$$



P(お金を やりとりした, D | EOS)
$\times$ P(彼女と, $P1$ | お金を やりとりした)

P(彼女とお金を やりとりした, D | EOS)

図 2: 生成確率計算の例

並列類似度は実数値をとるが、0 から 4 の整数値に離散化する。例えば、類似度が 1 未満の値をとるときは 0 に、4 以上の値をとるときには 4 になる。

例えば「彼女とお金をやりとりした」という文については、「彼女と」「お金を」間の並列類似度を 1 とすると、図 2 に示すように 2 つの依存構造候補が考えられる。T_2 は、格解析による生成確率 (後述) が高くなるため、T_1 よりも確率が高くなり選択されることになる。

節 C_i のタイプ (用言の活用や機能語の部分) を f_i とし、C_i から f_i を除いた内容部分を C'_i とする。C_{h_i} についても同様に、f_{h_i} と C'_{h_i} からなるとする。

$$\begin{aligned}&P(C_i, rel_{ih_i}|C_{h_i}) = P(C'_i, f_i, rel_{ih_i}|C'_{h_i}, f_{h_i}) \\ &= P(C'_i, rel_{ih_i}|f_i, C'_{h_i}, f_{h_i}) \times P(f_i|C'_{h_i}, f_{h_i}) \\ &\approx P(C'_i, rel_{ih_i}|f_i, C'_{h_i}) \times P(f_i|f_{h_i})\end{aligned}$$

この近似は、内容語は係り先のタイプには依存しない、またタイプは係り先には依存しないと考えられるからである。第 1 項を関係・内容部分生成確率、第 2 項をタイプ生成確率と呼ぶ。タイプ生成確率は、読点や提題助詞を生成する確率であり、[4] と同様に推定する。

関係・内容部分生成確率を以下のように変形、近似する。

$$\begin{aligned}&P(C'_i, rel_{ih_i}|f_i, C'_{h_i}) \\ &= P(C'_i|rel_{ih_i}, f_i, C'_{h_i}) \times P(rel_{ih_i}|f_i, C'_{h_i}) \\ &\approx P(C'_i|rel_{ih_i}, f_i, C'_{h_i}) \quad (\text{※内容部分生成確率}) \\ &\quad \times P(rel_{ih_i}|f_i) \quad (\text{※関係生成確率})\end{aligned}$$

最後の近似は、並列かどうかには、係り側のタイプ、さらには並列キーに特に依存していると考えられるためである。

以下では、まず関係生成確率について説明し、その後、内容部分生成確率について説明する。

3.1 関係生成確率

2つの句・節間が並列関係にあるかどうかには、並列キーがもっとも重要であると考えられる。そこで、関係生成確率の条件部のタイプ f_i を、並列キーをもつ場合には並列キー k_i、もたない場合には ϕ と考える。

関係生成確率は、並列キーをもつ場合には以下の確率となる。

$$P(rel_{ih_i}|f_i) = P(rel_{ih_i}|k_i)$$

並列キー k_i としては、約 50 種類の分類を用いる [6]。並列キーをもたない場合には、並列関係になることはないので、この確率は $P(D|\phi) = 1$ となる。関係生成確率の推定には、京都テキストコーパスを用いて最尤推定を行う。

3.2 内容部分生成確率

内容部分生成確率は、C'_i の主辞が述語かどうかによって場合分けを行う。

C'_i が述語を含む場合は、内容部分を述語項構造と考え、述語 v_i、格フレーム CF_l、格の対応関係 CA_k の 3 つからなると考える。格の対応関係 CA_k とは、入力側の格要素と格フレームの格との対応付けを表す。このときの内容部分生成確率 $P_p(C'_i|r_{ih_i}, f_i, C'_{h_i})$ は次のように書き換えられる。

$$\begin{aligned}
&P_p(C'_i|rel_{ih_i}, f_i, C'_{h_i}) \\
&= P(v_i, CF_l, CA_k|rel_{ih_i}, f_i, C'_{h_i}) \\
&= P(v_i|rel_{ih_i}, f_i, C'_{h_i}) \\
&\quad \times P(CF_l|rel_{ih_i}, f_i, C'_{h_i}, v_i) \\
&\quad \times P(CA_k|rel_{ih_i}, f_i, C'_{h_i}, v_i, CF_l) \\
&\approx P(v_i|rel_{ih_i}, f_i, w_{h_i}) \quad (\text{※内容語生成確率}) \\
&\quad \times P(CF_l|v_i) \qquad (\text{※格フレーム生成確率}) \\
&\quad \times P(CA_k|CF_l, f_i) \quad (\text{※格の対応関係生成確率})
\end{aligned}$$

この近似は、述語 v_i は係り先の節の中では主辞単語 w_{h_i} に、格フレーム CF_l は述語 v_i のみに、格の対応関係 CA_k は格フレーム CF_l と付属語列 f_i に依存すると考えられることによるものである。

内容部分生成確率を構成する確率の中で、我々がすでに提案している構文・格構造解析の統合的確率モデル [3] と異なるのは内容語生成確率のみである。以下ではこの確率について説明する。

表 1: 並列構造の検出の精度

	ベースライン	提案手法
適合率	366/460 (79.6%)	361/435 (83.0%)
再現率	366/447 (81.9%)	361/447 (80.8%)
F 値	– (80.7%)	– (81.9%)

内容語生成確率の条件部のタイプ f_i は、関係生成確率と同様に、並列キーに依存すると考えられるので、内容語生成確率を次のように定義する。

$$P(v_i|rel_{ih_i}, f_i, w_{h_i}) = P(v_i|rel_{ih_i}, k_i, w_{h_i})$$

ただし、並列ではない場合 ($rel_{ih_i} = D$) には、並列キーは考慮しないこととする。

C'_i は、述語を含まない場合は、連体修飾句か並列前部の句のどちらかであり、名詞 n_i を生成する確率と考える。現在は名詞格フレームを用いていないため、内容部分生成確率は内容語生成確率のみからなり、以下のような確率となる。

$$P_n(C'_i|rel_{ih_i}, f_i, C'_{h_i}) \approx P(n_i|rel_{ih_i}, f_i, w_{h_i})$$

内容語生成確率は、関係付きの単語間の共起から推定できる。この推定は、大規模コーパスの自動解析結果を利用して最尤推定により行う。

4 実験

提案手法によって解析した並列・構文構造の評価実験を行った。格フレームはウェブテキスト約 5 億文から自動構築したものを用い、構文・格解析済みデータはウェブテキスト約 1 億文を構文・格解析することによって得たものを用いた。このときに用いた格解析は、シソーラスに基づく類似度を用いた格解析 [2] である。

本実験は、ウェブテキスト 759 文*を形態素解析器 JUMAN†に通した結果を提案システムに入力することによって行う。その 759 文には、京都テキストコーパスと同じ基準でタグ付けを行っており、これを用いて評価を行う。

ベースラインとしては、構文解析器 KNP‡と、確率的構文・格解析 [4] の 2 つを用いた。両方のシステムとも並列構造解析は、構文解析を行う前に、2 節の類似度スコアが閾値を越えて最大となる並列構造に一意に決定する手法を用いている。

*これらの文は格フレーム構築とモデル学習には用いていない。
†http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
‡http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html

表 2: 構文構造の精度

構文解析	構文・格解析	提案手法
3833/4436 (86.4%)	3852/4436 (86.8%)	3893/4436 (87.8%)

4.1 並列構造検出の評価

並列構造検出の評価は、並列キーの文節の係り関係が並列 (P) である場合について、解析結果と正解を比較することによって行った。その結果を表 1 に示す。提案手法は、ベースライン (KNP) より F 値で 1.2%改善された。

4.2 構文構造の評価

構文構造の評価には、並列構造の範囲同定の評価も含むものと考える。構文構造の評価としては、文末から 2 つ目までの文節以外の係り受けを評価する。その評価結果を表 2 に示す。表において、「構文解析」とは構文解析器 KNP による精度であり、「構文+格解析」は確率的構文・格解析 [4] による精度である。提案手法の精度は、「構文解析」「構文+格解析」のそれぞれに対して 1.4%、1.0%向上した。どちらの精度向上もマクネマー検定の結果、有意 ($p < 0.01$) であった。提案手法は、述語項構造まわりの関係だけではなく、体言間、用言間の並列関係の解析に有効であった。

表 3 に、「構文+格解析」では誤りになるが、提案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。例えば (1) の例では、「アプリケーション」と「ドライバ」が並列であることが正しく解析できるようになった。これは、「ドライバ」が「(プレ) インストールされている」の格フレームから生成されやすいことと、「アプリケーション」と「ドライバ」が並列になりやすいことが考慮されたためである。

4.3 議論

構文解析誤りの原因としては、提案手法が強く語彙的選好を考慮しているため、係り受けの正解基準からずれることによるものがある。

- 行政相談委員は、 いつでも自宅でみなさんからのご相談に 応じていますが、× この期間中は次のところで行政相談所を 開きます。○

この文において、「行政相談委員は、」の正解係り先は「開きます。」であるが、提案手法は係り先を「応じていますが、」と解析し、誤りとなる。「開きます。」、「応じていますが、」のどちらも意味的には係り先として正しいと考えられるが、基準としては文末の「開きます。」であるのでずれが生じる。このような問題を解決するには、省略関係の正解を考慮しながら評価を行う必要がある。また、このような正解基準からのずれによる誤りは並列構造検出についても生じ、特に述語並列の場合に正解を並列構造とするかどうかで揺れるケースがある。

表 3: 解析が正しくなった例

(1) ＩＥＥＥ１３９４標準搭載のパソコンでは、プレインストールされている アプリケーションおよびドライバ○ との競合により動作しない 場合が× あります。

(2)産業資源部と韓国貿易協会によると、昨年１月から１２月末までの対中南米貿易収支黒字は ４１億４２００万ドルと、 前年より× １０億ドル 減少した。○

(3)相談がまとまった グループは すぐ 動き出すので○ わかるのだが、なかなか動き出さないグループも ある。×

本研究で提案した生成的確率モデルは、言語モデルとして使えるというような利点があるが、並列構造や構文・格構造の同定に有効な素性を自由に選択できないという問題がある。今後、この問題に対処するためには、識別モデルを導入することが考えられる。

参考文献

[1] Rajeev Agarwal and Lois Boggess. A simple but useful approach to conjunct identification. In *Proceedings of ACL1992*, pp. 15–21, 1992.

[2] Daisuke Kawahara and Sadao Kurohashi. Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of COLING2002*, pp. 425–431, 2002.

[3] 河原大輔, 黒橋禎夫. Web から獲得した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. 言語処理学会 第 12 回年次大会, pp. 1111–1114, 2006.

[4] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会 自然言語処理研究会 2006-NL-171, pp. 67–73, 2006.

[5] 国立国語研究所. 分類語彙表 増補改訂版. 大日本図書, 2004.

[6] 黒橋禎夫, 長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol. 1, No. 1, pp. 35–57, 1994.