Vol.2017-NL-231 No.19
Vol.2017-SLP-116 No.19
2017/5/16

ニューラル言語モデルを用いた 法令文の並列構造解析とその評価

山腰 貴大[1,a)]　大野 誠寛[1,2,3,b)]　小川 泰弘[1,2,3,c)]　中村 誠[4,d)]　外山 勝彦[1,2,3,e)]

概要：法令文は，一般の人々にとって読みにくいものであるとされている．その原因の一つは，階層的な並列構造が多用されることであると考えられる．複雑な階層を持つ並列構造は，人間による可読性を低下させるだけでなく，機械による法令文書処理の性能を低下させる要因にもなる．そのため，法令文書処理において，高性能な並列構造解析技術の確立が望まれる．法令文に対する並列構造解析手法はすでに提案されているが，十分な解析性能を達成しているとは言い難い．その原因の一つに，並列構造の同定の手がかりとして，単語アライメントに基づく句の類似度を用いていることが挙げられる．そこで本稿は，ニューラル言語モデル (NLM) を用いた法令文の新しい並列構造解析手法を提案する．提案手法は，文脈を考慮した並列句間の類似性や，並列句を互いに入れ替えたときの文の流暢性を NLM によって求め，それらに基づいて並列構造を決定的に同定する．評価実験の結果，提案手法の F 値は 64%となり，従来手法と比べて解析性能が 25%程向上した．

キーワード：ニューラル言語モデル，法令文，並列構造解析，法令文書処理

1. はじめに

法令文は，一般の人々にとって読みにくいものであるとされている．その原因の一つは，階層的な並列構造が多用されることであると考えられる．例えば，図 1 の法令文には，階層を成した 4 個の並列構造が含まれている．このように複雑な階層を持つ並列構造は，人間による可読性を低下させるだけでなく，機械による法令文書処理の性能を低下させる要因にもなる．そのため，法令文の読解支援 [1] や法令用語シソーラスの自動構築 [2] などの法令文書処理においては，法令文に対する高性能な並列構造解析技術が望まれている．

並列構造解析に関する研究は，一般文を対象として，これまでに数多く行われている（例えば，[3][4]）．しかし，これらの手法は法令文特有の性質を考慮していないため，法令文に対する十分な解析が行えないと考えられる．一方で，法令文を対象とした並列構造解析手法として，松山らの手法 [5]（以下，従来手法）がある．この手法は，法令文における等位接続詞の使い分け [6][7] に基づいて並列構造を決定的に同定する．しかし，従来手法は十分な解析性能を達成しているとは言い難い．その原因の一つとして，並列構造の同定の手がかりに，一対一の単語アライメントに基づく句の類似度を用いていることが挙げられる．すなわち，従来手法は，単語数の異なる句が並列関係にある場合，それらの間の類似度を過少に算出するため，その同定に失敗する傾向にある．

そこで，本稿では，ニューラル言語モデル（NLM）[8] を用いた法令文の並列構造解析手法（以下，提案手法）を提案する．提案手法は，文脈を考慮した句の間の類似性や，並列関係にある句を互いに入れ替えたときの文の流暢性を NLM によって求め，それらに基づいて並列構造を決定的に同定する．並列関係にあると想定される句を一つのベクトルとして扱うため，句の内部の単語数に依存せずに並列構造を同定できる．

提案手法の有効性を検証するために，構文情報付きの法令文コーパス [9] に収録された法令を対象に並列構造解析を行った．その結果，提案手法は精度・再現率ともに従来手法を上回った．

本稿の構成は次の通りである．次の 2 節で法令文に特有な並列構造について解説し，3 節で従来手法の概要を説明

[1] 名古屋大学 大学院情報科学研究科
[2] 名古屋大学 大学院情報学研究科
[3] 名古屋大学 情報基盤センター
[4] 名古屋大学 大学院法学研究科
[a)] yamakoshi@kl.i.is.nagoya-u.ac.jp
[b)] ohno@nagoya-u.jp
[c)] yasuhiro@is.nagoya-u.ac.jp
[d)] mnakamur@nagoya-u.jp
[e)] toyama@is.nagoya-u.ac.jp

Vol.2017-NL-231 No.19
Vol.2017-SLP-116 No.19
2017/5/16

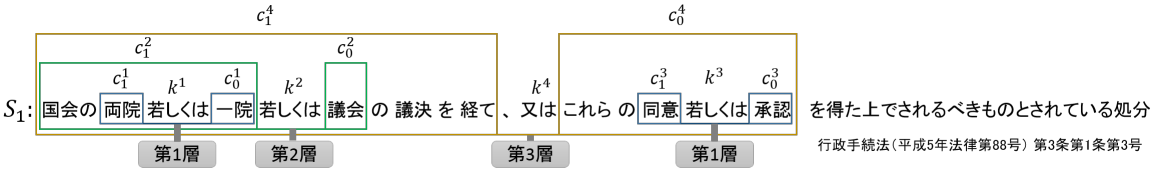

図 1　階層的並列構造を含む法令文の例

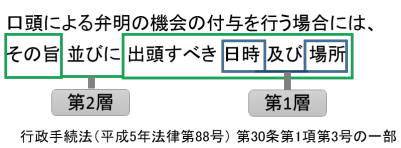

図 2　「及び」と「並びに」による階層的並列構造の例

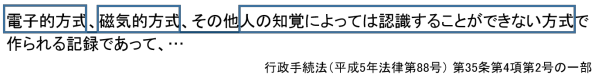

図 3　「その他」による並列構造の例

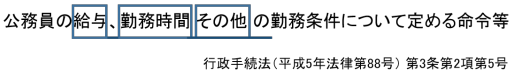

図 4　「その他の」による並列構造の例

する．4 節で NLM の基本的な概念を解説した後，5 節で提案手法について，従来手法との相違点を中心に説明する．6 節で評価実験とその考察について述べ，7 節でまとめと今後の課題を述べる．

2. 法令文に特有な並列構造

本節では，構文情報付き法令文コーパス作成のための基準 [9] に基づいて，法令文特有の並列構造を簡単に解説する．なお，本稿では，図 1 の「両院若しくは一院」における「両院」や「一院」のように，並列関係にある語句を並列句と呼ぶ．また，「若しくは」，「又は」のように，並列句の連接と並列の種類を明示する語句を並列キーと呼ぶ．

2.1 階層的並列構造

法令文において階層的並列構造を表すための並列キー「又は」と「若しくは」，「及び」と「並びに」は，それぞれ使い分けられる [6][7]．

「又は」と「若しくは」は選択的な並列を表す．一般文において，これらの並列キーを意識して使い分けることはない．しかし，法令文においては，「又は」は最上位の並列構造に対して，「若しくは」はそれ以外の階層の並列構造に対して，それぞれ用いられる．図 1 の文 S_1 中に現れる並列キーは，この規則に基づいて使い分けられている．

一方，「及び」と「並びに」は併合的な並列を表す．「又は」と「若しくは」同様に，一般文ではこれらの並列キーを意識して使い分けることはない．しかし，法令文においては，「及び」は最下位の並列構造に対して，「並びに」はそれ以外の階層の並列構造に対して，それぞれ用いられる．「及び」と「並びに」による階層的並列構造の例を図 2 に示す．

2.2 「その他」と「その他の」による並列構造

「その他」と「その他の」は語句を例示的に列挙する場合に用いられる．一般文において，これら二つは区別されることなく用いられる．しかし，法令文では明確に区別され，それぞれ異なる意味を持つ [6][7]．

「その他」は，事物を並列的に例示する際に用いられる．「その他」による並列構造の例を図 3 に示す．図 3 において，「電子的方式」，「磁気的方式」，「人の知覚によっては認識することができない方式」がそれぞれ並列関係にある．この例における「人の知覚によっては認識することができない方式」は，一般に「人の知覚によっては認識することができない方式」とされるものの集合から「電子的方式」と「磁気的方式」を除いたものを指し示している．このように，法令文において「その他」の後方にある語句は，前方にある語句と並列関係にあり，その上で，ある事物の集合から前方の語句が指し示すものを除いたものを表す [9]．

一方，「その他の」は，その前方に挙げられた事物がその後方に挙げられた事物の下位概念であることを示す際に用いられる．「その他の」による並列構造の例を図 4 に示す．図 4 において，「給与」と「勤務時間」と「その他」は互いに並列関係にあるが，これらの上位概念を示す「勤務条件」とは並列関係にない．ここで，この場合の「その他」は，「勤務条件」から「給与」と「勤務時間」を除いたものを指し示しており，「給与」や「勤務時間」と意味上同等であると考えられ，これらと並列関係にある．一方，「勤務条件」には「給与」「勤務時間」以外の事物も含まれ，「給与、勤務時間その他」と「勤務条件」はそれぞれ同じ事物を指し示し，意味を限定し合っていることから，「給与、勤務時間その他」と「勤務条件」は並列関係ではなく同格関係にある [9]．したがって，法令文において「その他の」の前方にある語句と，「その他の」の中の「その他」とは互いに並列関係にある一方で，これらは，後方にある語句と並列関係にはない．

3. 従来手法

提案手法は，従来手法 [5] をもとにして，並列句と並列構造を決定的に解析する手順は踏襲し，解析性能を低下させる原因と考えられる部分を改良することにより，高い解

Vol.2017-NL-231 No.19
Vol.2017-SLP-116 No.19
2017/5/16

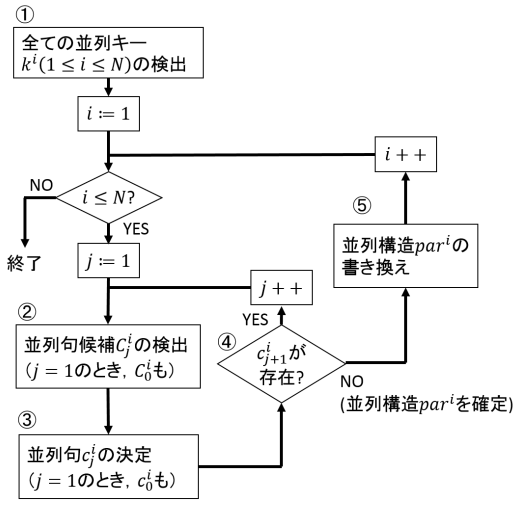

図 5 並列構造解析処理の流れ

$$\left(\begin{array}{c}\text{及び}\\ \text{若しくは}\end{array}\right) > \left(\begin{array}{c}\text{並びに}\\ \text{又は}\end{array}\right) > \left(\begin{array}{c}\text{その他}\\ \text{かつ}\\ \text{と}\\ \text{や}\end{array}\right)$$

図 6 従来手法における並列キーとその優先順位

析性能を備えた並列構造解析を実現する．そこで，本節では，従来手法について，解析性能低下の原因と考えられる点を述べつつ，その解析手順を説明する．

従来手法は，図 5 に示す手順により，1 文中のすべての並列構造 par^i $(1 \le i \le N)$ を順に同定する．ここで，N は文中に出現する並列キーの数である．この手法では，一つの並列構造 par^i の単語列は次の式 (1) で形式化できることを前提とし，その内部の並列句 c_j^i $(j \ge 0)$ を文末から文頭に向かって順に決定することにより，並列構造 par^i を同定する．

$$par^i = (c_j^i \cdot \text{“、”})^* \cdot c_1^i \cdot k^i \cdot c_0^i \qquad (1)$$

ここで，k^i は par^i を構成する並列キーを，“·” は語の連接をそれぞれ表す．図 1 における記号 c, k は，式 (1) に基づいて各並列構造の各要素を表す．以下の 3.1 節から 3.5 節では，図 5 の処理①から処理⑤をそれぞれ詳述する．

3.1 並列キーの検出

図 6 に示す並列キーを対象として，1 文中のすべての並列キー k^i $(1 \le i \le N)$ を検出する．ここで，i は，図 6 の優先順位（同順位の場合は，文頭に近い順）に基づいて並列キーを並び替えた後の順位を示す．

3.2 並列句候補の検出

並列句 c_j^i の候補の集合 C_j^i を求める．j の値が 0 か，1 か，2 以上かによって候補の求め方は異なるが，候補となる単語列が読点や未処理の並列キーを含まないことを共通の条件とする．また，同定済みの並列構造を含める場合，

行政庁は、申請をしようとする者又は申請者の求めに応じ、...
行政手続法（平成5年法律第88号）第9条第2項の一部

図 7 従来手法が解析に失敗しやすい並列構造の例

図 8 （左）「申請をしようとする者」と「申請者」の単語アライメント
（右）「者」と「申請者」の単語アライメント

その全体を必ず含めることとする．

C_1^i に含まれる単語列の始点は文節の最左語で，終点は必ず並列キー k^i の直前の語（読点の場合はその前の語）である．

C_0^i に含まれる単語列の始点は必ず k^i の直後の語（読点の場合はその後の語）で，終点は c_1^i の各候補の最右語と同じ品詞で文節末に最も近い語である．c_1^i の候補の最右語が名詞であるとき，c_1^i の候補の最右語と c_0^i の候補の最右語でシソーラスに基づく語の意味的類似性を計算し，類似度が高い上位 3 候補を残す [*1]．

C_j^i $(j \ge 2)$ に含まれる単語列の始点は文節の最左語で，終点は c_{j-1}^i の始点の二つ前の語である [*2]．

例えば，図 1 の文 S_1 において，並列キー k^1 に対する候補集合 C_1^1, C_0^1 は，それぞれ { 国会の両院, 両院 }, { 一院 } となる．

3.3 並列句の決定

c_1^i と c_0^i は同時に決定する．具体的には，c_1^i の候補と c_0^i の候補のすべての組み合わせから，並列句間の類似度が最も高くなるものを選択する．一方，c_j^i $(j \ge 2)$ に関しては，c_j^i の各候補と決定済みの c_{j-1}^i との間で類似度を計算し，類似度が最も高くなる候補を c_j^i として決定する．

並列句間の類似度は，並列句候補ペアに対する一対一の単語アライメントを考え，対応関係を持つ単語の割合と，対応関係にある単語間の類似度の和により求める．単語間の類似度は，二つの単語が同じ品詞の場合に高くなる．また，対象となる単語が名詞の場合は，意味的類似性の値も加味される．単語アライメントを一対一に制限することにより，最適な単語アライメントを動的計画法によって計算することが可能となる．

しかし，一対一の単語アライメントを用いた計算方法は，単語数が大きく異なる並列句の同定を苦手とする．図 7 に従来手法が解析に失敗しやすい並列構造の例を示す．「申請をしようとする者」と「申請者」の単語アライメントは図 8 の左図となる．一つの単語は複数の単語と対応関係を

*1 同定済みの並列構造の終点は類似度に関わらず候補として残す．
*2 c_{j-1}^i の始点の一つ前の語は読点であるため．

持てないため，単語「を」，「し」，「よう」，「と」，「する」はいずれも対応関係を持たないことになる．従来手法の類似度計算では，これらの対応関係を持たない単語の数に応じてペナルティを設定している．そのため，単語数が大きく異なる句の間の類似度は低く算出される傾向にある．一方，「者」と「申請者」の単語アライメント（図 8 の右図）では，互いの単語数が近いため，対応関係を持たない単語が減少し，類似度計算におけるペナルティが減る．結果的に，従来手法では，「者」と「申請者」の類似度を最も高く算出し，図 7 の並列構造の同定に失敗する．

3.4 さらなる並列句の存在判定

並列句 c^i_j の直前の語が読点であり，かつ，その前の語と c^i_j の最右語が同じ品詞の場合， c^i_{j+1} が存在すると判定する．ただし，その品詞が名詞の場合，二つの語の意味的類似性を計算し，閾値未満であった場合は c^i_{j+1} が存在しないと判定する．例えば，図 1 の文 S_1 において， c^1_1 を「両院」と決定した後の c^1_2 の存在判定を考える．このとき，「一院」の直前の語「の」は読点でないため， c^1_2 は存在しないと判定する．

3.5 並列構造の書き換え

並列句のどちらか一方の内部に下位の並列構造が存在すると，それらの間の類似度が低く算出される．これを避けるため，同定した並列構造 par^i の全体を c^i_0 で書き換える．

図 1 の文 S_1 において，c^1_1 を「両院」，c^1_0 を「一院」と決定し，並列構造 par^1 の同定処理が終わった後，S_1 を「国会の一院若しくは議会の…」に書き換える．

4. ニューラル言語モデル

提案手法は，3.3 節で述べた従来手法の類似度計算における問題を解消し，高い解析性能を実現するために，ニューラル言語モデル（NLM）を利用する．本節では，提案手法の詳細な説明に先立って，NLM の基本的な概念を説明する．

NLM[8] は，ニューラルネットワークを用いた言語モデルである．入力された語をベクトルに変換し，次に現れる語の確率分布を出力する．最新の NLM の多くは，再帰型ニューラルネットワーク（RNN）[10] によって構築されている．RNN は再帰的な結合を持つため，過去の情報，すなわち文脈を利用して出力値を計算できる．

図 9 に，NLM を構成する RNN の例を示す．この例では，再帰的結合を持つ二つの隠れ層 h_1 と h_2 が文脈情報を保持している．なお，図 9 は，6 節で述べる評価実験において提案手法が用いた RNN の構成である．

5. 提案手法

提案手法は，形態素情報と文節境界情報が付与された 1 文を入力とし，1 文中に存在するすべての並列構造を出力とする．従来手法と同じ手順（図 5）で 1 文に対する並列構造解析を進めるが，並列句の決定において NLM を用いる点に大きな特徴がある．本節では，図 5 の処理①から処理⑤について，従来手法と異なる点を中心に述べる．

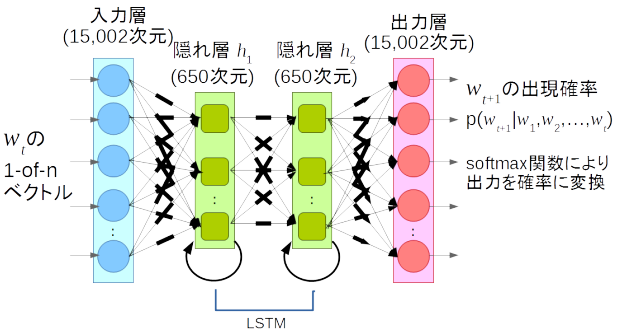

図 9 NLM を構築する RNN の例（提案手法で用いた RNN）

(から) > (及び / 若しくは) > (並びに / 又は) > (その他 / その他の / かつ)

図 10 提案手法における並列キーとその優先順位

同項第九号から第四号までに掲げるものにあっては、
技術上の秘密に関するものに限る。
不正競争防止法（平成5年法律第47号）第5条第1項

図 11 「から」による並列構造の例

5.1 並列キーの検出

図 10 の並列キーを対象として，入力文中のすべての並列キーを検出した後，その優先順位に基づき番号付けを行う．提案手法では，並列キー「その他の」は並列キー「その他」と用法が異なるため，また，並列キー「から」は条番号などの並列（図 11 参照）において頻繁に使用されるため，これら二つの並列キーをそれぞれ解析の対象に加えることとした．二つの並列キー「その他の」と「から」により構成される並列構造は，それぞれ次の式 (2) と式 (3) で形式化されるものであり，3 節で示した式 (1) とは異なる．そのため，これらの並列構造の同定の際に一部例外的な処理を施す．

$$par^i_{\text{その他の}} = (c^i_j \cdot \text{“、”})^* \cdot c^i_1 \cdot \text{“その他の”}, \quad (2)$$

$$par^i_{\text{から}} = c^i_1 \cdot \text{“から”} \cdot c^i_0 \cdot \text{“まで”} \quad (3)$$

なお，提案手法では，図 6 の並列キーのうち，「や」や「と」を解析の対象から外した．6 節で述べる実験データ [*3] において，「や」はまったく出現せず，「と」も並列キー全体のわずか 0.70% (5/717) しか出現しなかったためである．

[*3] 不正競争防止法，行政手続法より作成．

5.2 並列句候補の検出

並列句 c_j^i の候補の集合 C_j^i は，原則的に従来手法と同様に求める．ただし，提案手法はシソーラスを用いないため，これに由来する制約は設けない．一方，法令文における並列キーの使用規則 [6][7] に忠実に従うため，従来手法では設けられていなかった次の制約を追加する．

- 「及び」による並列構造が「及び」による別の並列構造を下位に持つことは禁止されているため，並列キー k^i が「及び」のとき，「及び」による同定済みの並列構造を含まないようにする．
- 「又は」による並列構造が「又は」による別の並列構造を下位に持つことは禁止されているため，並列キー k^i が「又は」のとき，「又は」による同定済みの並列構造を含まないようにする．
- 「若しくは」による並列構造は必ず「又は」による並列構造を上位に持つため，並列キー k^i が「又は」のとき，他の制約に違反せずに含められる場合，「若しくは」による同定済みの並列構造を必ず含める．

例外的に，並列キー k^i が「その他の」または「から」のときは，最右の並列句の候補集合 C_0^i を次の通りに求める．

- k^i が「その他の」のとき，C_0^i の要素は k^i 内の「その他」一つのみとする．
- k^i が「から」のとき，C_0^i の要素は一つのみとし，その要素は k^i の直後の語から k^i と最も近い「まで」の直前の語までの語句とする．ただし，「まで」が存在しない，あるいは k^i から当該「まで」の間に読点が存在する場合，k^i は並列構造を構成しないと考える．

5.3 並列句の決定

提案手法は，「並列句は互いに類似し，かつ，並列句の順序を入れ替えても文の流暢性は保たれる」という仮定に基づいて並列句を決定する．具体的には，入力文を S とするとき，並列句 c_1^i, c_0^i は次の式 (4) により同時に決定し，並列句 c_j^i $(j \geq 2)$ は式 (5) により決定する．

$$(c_1^i, c_0^i) = \underset{(c_l, c_r) \in C_1^i \times C_0^i}{\arg\max} \mathit{sim}(S, (c_l, c_r), c_r) \times \mathit{flu}(S, (c_l, c_r)), \tag{4}$$

$$c_j^i = \underset{c_l \in C_j^i}{\arg\max}\, \mathit{sim}(S, (c_l, c_{j-1}^i), c_0^i) \times \mathit{flu}(S, (c_l, c_{j-1}^i)) \tag{5}$$

ここで，$\mathit{sim}(S, (c_l, c_r), c_0)$ は並列句の類似性を表すスコア（類似性スコア）であり，$\mathit{flu}(S, (c_l, c_r))$ は並列句の順序を入れ替えたときの文の流暢性を表すスコア（流暢性スコア）である．どちらのスコアも，通常の語順の NLM（F-NLM）と，語順を逆にした NLM（B-NLM）によって計算される．

5.3.1 類似性スコア

類似性スコア $\mathit{sim}(S, (c_l, c_r), c_0)$ は次の式 (6) により求める．

$$\begin{aligned} &\mathit{sim}(S, (c_l, c_r), c_0) \\ &\quad = \mathit{sim}_f(S, (c_l, c_r), c_0) + \mathit{sim}_b(S, (c_l, c_r), c_0), \end{aligned} \tag{6}$$

$$\mathit{sim}_f = 1 + |\mathit{sim}_c(\mathit{vec}_f(W_{fl}), \mathit{vec}_f(W_{fr}))|, \tag{7}$$

$$\mathit{sim}_b = 1 + |\mathit{sim}_c(\mathit{vec}_b(W_{bl}), \mathit{vec}_b(W_{br}))| \tag{8}$$

ここで，$\mathit{sim}_c(\boldsymbol{u}, \boldsymbol{v})$ は，二つのベクトル $\boldsymbol{u}$ と $\boldsymbol{v}$ のコサイン類似度である．予備実験により，コサイン類似度の絶対値に 1 を加えたものを各言語モデルによる類似性スコアとした．

一方，$\mathit{vec}_f(W)$ は F-NLM に単語列 W を入力し終えたときの，また，$\mathit{vec}_b(W)$ は B-NLM に W を逆順にしたものを入力し終えたときの隠れ層 h_2 の値をそれぞれ表す．隠れ層の値を用いることにより，文脈を考慮した類似性を捉えられると期待できる．

式 (7), 式 (8) 中の単語列 $W_{fl}, W_{fr}, W_{bl}, W_{br}$ は，それぞれ次の式 (9) から式 (12) により生成する．

$$W_{fl} = W_f \cdot c_l, \tag{9}$$

$$W_{fr} = W_f \cdot c_r, \tag{10}$$

$$W_{bl} = c_l \cdot W_b, \tag{11}$$

$$W_{br} = c_r \cdot W_b \tag{12}$$

ここで，W_f は並列句 c_l の前方にある単語列，また，W_b は並列句 c_0 の後方にある単語列をそれぞれ表す．並列構造の前後にある単語列も用いることにより，文脈をより詳細に捉えられることを期待できる．

例えば図 1 において，$\mathit{sim}(S_1, (「\text{両院}」, 「\text{一院}」), 「\text{一院}」)$ を計算するときの W_{fl}, W_{fr}, W_{bl}, W_{br} は，それぞれ式 (13) から式 (16) の通りになる．

$$W_{fl} = 「\text{国会の}\underline{\text{両院}}」, \tag{13}$$

$$W_{fr} = 「\text{国会の}\underline{\text{一院}}」, \tag{14}$$

$$W_{bl} = 「\underline{\text{両院}}\text{若しくは議会の…}」, \tag{15}$$

$$W_{br} = 「\underline{\text{一院}}\text{若しくは議会の…}」 \tag{16}$$

5.3.2 流暢性スコア

流暢性スコア $\mathit{flu}(S, (c_l, c_r))$ は式 (17) により求める．

$$\mathit{flu}(S, (c_l, c_r)) = \mathit{flu}_f(W_s) + \mathit{flu}_b(W_s) \tag{17}$$

ここで，W_s は，S 中の二つの並列句 c_l と c_r を入れ替えた単語列とする．例えば，図 1 の文 S_1 において，$\mathit{flu}(S_1, (「\text{両院}」, 「\text{一院}」))$ を計算するときの W_s は「国会の$\underline{\text{一院}}$若しくは$\underline{\text{両院}}$若しくは議会の…」となる．$\mathit{flu}_f(W_s)$ と $\mathit{flu}_b(W_s)$ は，それぞれ F-NLM と B-NLM に基づく W_s の流暢性を返す関数であり，次の式 (18) と式 (19) により求める．

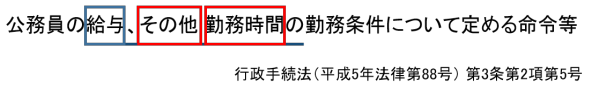

図 12 「その他の」の並列構造の入れ替え

$$flu_f(W_s) = \sqrt[|W_s|]{\prod_{t=1}^{|W_s|} P_f(w_t|w_1, w_2, \ldots, w_{t-1})}, \quad (18)$$

$$flu_b(W_s) = \sqrt[|W_s|]{\prod_{t=1}^{|W_s|} P_b(w_t|w_{|W_s|}, w_{|W_s|-1}, \ldots, w_{t+1})} \quad (19)$$

ここで，$P_f(w_t|w_1, w_2, \ldots, w_{t-1})$ は，F-NLM において，単語列 w_1, w_2, ..., w_{t-1} の次に w_t が現れる確率を表す．$P_b(w_t|w_{|W_s|}, w_{|W_s|-1}, \ldots, w_{t+1})$ は，B-NLM における同様の確率を表す．文長の影響を排除するために，確率の相乗平均を求める．

例外として，並列キー k^i が「から」のときは，流暢性スコアを求めない．「から」（〜まで）は条番号などの並列に用いられ，通常，並列句は条番号などの昇順に配置される．本節における並列句の入れ替えを行うと逆順になるが，そのように記述されることはないため，提案手法が流暢性を導入した意図に反して，スコアが大きく低下する恐れがある．

また，k^i が「その他の」であり，並列句 c_i^1, c_i^0 を求める場合も流暢性スコアを求めない．k^i が「その他の」のとき，c_0^i は「その他」となるが，この語句と c_1^i を入れ替えると日本語として不自然である文，あるいは，読みにくい文となり，流暢性スコアが低く算出されることになる．例えば，図 4 における二つの並列句「勤務時間」と「その他」の順序を入れ替えると図 12 に示す文となり，日本語として不自然な文となることが分かる．

5.4 さらなる並列句の存在判定

並列句 c_j^i の前方に c_{j+1}^i が存在するかどうかの判定は，原則的に従来手法と同様に行う．ただし，提案手法はシソーラスを用いないため，意味的類似性に関する制約を設けない．

5.5 並列構造の書き換え

従来手法と同様に，同定した並列構造 par^i 全体を c_0^i で書き換える．

6. 評価実験

提案手法の有効性を検証するため，構文情報付きの法令文コーパス [9] に収録された 2 法令（不正競争防止法，行政手続法）を対象に並列構造解析を行った．

6.1 実験概要

上記の法令文コーパス [9] から入力データと正解データを作成した．コーパスは 1 文で 1 ファイルとなっており，形態素情報，文節境界情報，係り受け情報が付与されている．コーパスから係り受け情報を取り除き，括弧で囲まれた文字列は別の文として独立させたものを入力データとして使用した．入力データは合計 592 文で，717 個の並列構造が存在する．正解データは，コーパスの係り受け情報から作成した．

性能比較のため，従来手法を独自に実装したシステムにより，同じ法令文に対する並列構造解析を行った．システムのパラメータは，松山らの論文 [5] で示されている値を用いた．

NLM 学習用コーパスは，日本法令外国語訳データベースシステム（JLT）[*4] より 2016 年 9 月にダウンロードした法令文（574,062 文）から作成した．

NLM は，図 9 に示す RNN により構築し，単語の基本形を RNN の入力とした．隠れ層の数は，予備実験の結果により 2 層とした．形態素解析は MeCab (v0.98) [11] で行い，辞書は IPA 辞書を用いた．出現頻度が高い 15,000 語と終端記号，未知語を有効語彙としたため，入力層と出力層は 15,002 次元である．

NLM の学習は，Chainer (v1.15.0) [*5] を介して行った．パラメータの更新は確率的勾配降下法（学習率 1）により行い，更新時に，勾配ベクトルの L2 ノルムの最大値を 5 に設定し，ユニットを 0.5 の確率でドロップアウトさせた．エポック数は 8 とした．

評価では，正しく同定した並列構造と並列句の精度・再現率をそれぞれ求めた．精度は，解析結果中の並列構造や並列句において正しく同定したものの割合であり，また，再現率は，正解データ中の並列構造や並列句において正しく同定したものの割合である．解析結果中の並列句の範囲が正解データのものと完全に一致した場合，その並列句を正しく同定したと判定した．また，解析結果中の並列構造において，すべての並列句の範囲が正解データと完全に一致した場合，その並列構造を正しく同定したと判定した．なお，最右の並列句の候補を一つも検出できなかった（C_0^i が空となった）場合，並列構造 par^i の解析は完遂されず，その解析結果は出力されなかったものと考え，精度計算の分母には含めていない．

6.2 実験結果

表 1 に，各手法で並列構造解析を行った結果を示す．提案手法は，従来手法と比べて，並列構造単位・並列句単位のどちらにおいても精度・再現率ともに大幅に向上した．

並列構造の解析に成功した例を図 13 に示す．従来手法

*4 http://www.japaneselawtranslation.go.jp/
*5 http://chainer.org/

Vol.2017-NL-231 No.19
Vol.2017-SLP-116 No.19
2017/5/16

表 1　実験結果

		提案手法	従来手法
並列構造	精度	66.1%(449/679)	36.1%(286/793)
	再現率	62.6%(449/717)	39.9%(286/717)
	F 値	64.3%	37.9%
並列句	精度	83.6%(1,339/1,602)	55.9%(976/1,747)
	再現率	79.0%(1,339/1,694)	57.6%(976/1,694)
	F 値	81.2%	56.7%

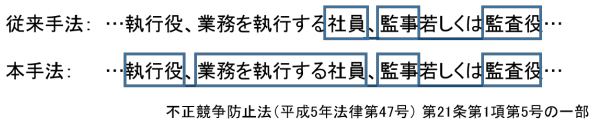

図 13　解析に成功した例

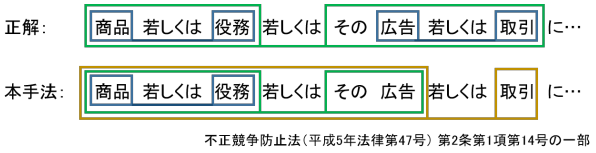

図 14　解析に失敗した例

は並列句「業務を執行する社員」の決定に失敗し，さらに前方にある並列句「執行役」を探索できなかったが，提案手法は並列構造を正しく解析できた．従来手法は，比較する並列句候補ペアに対して，一対一の単語アライメントを考え，対応関係を持つ単語の割合と対応関係にある単語間の類似度の和に基づいて，並列句間の類似度を構成的に計算する．そのため，並列句内の単語構成に影響を受けやすく，特に，単語数が大きく異なる場合，並列句間の類似度を過少に算出してしまう．その結果として，図 13 のように単語数が近い候補「社員」を並列句として選択してしまったと考えられる．一方，提案手法は，並列句候補を（その前後の単語列を考慮しつつ）NLM によって一つのベクトルに変換し，それを用いて並列句候補間の類似度を直接的に計算するため，並列句候補の単語数の影響を受けることなく，正しく解析したと考えられる．

次に，解析に失敗した例を図 14 に示す．図 14 の法令文中には，並列キー「若しくは」が三つ現れ，二番目の「若しくは」に対応する並列構造が最上位となる．しかし，提案手法は，文頭に近い並列キーに対応する並列構造から順番に同定するため，図 14 のように誤った解析を必ず行う．なお，従来手法においても，提案手法と同じ順序で並列構造を同定するため，同様に解析に失敗する．

6.3　考察

本節では，提案手法の特徴や有効性をより詳細に明らかにするために，並列構造を構成する並列キーの種類，並列構造が属する階層，並列構造が持つ並列句の数にそれぞれ着目して解析結果を分析する．

表 2　並列キーの種類ごとの評価結果
（下線は提案手法で対象とした並列キー）

並列キー	提案手法		従来手法	
	精度	再現率	精度	再現率
又は	65.0% (165/254)	62.7% (165/263)	47.2% (119/252)	45.2% (119/263)
若しくは	69.1% (125/181)	69.1% (125/181)	60.9% (106/174)	58.6% (106/181)
及び	77.2% (88/114)	77.2% (88/114)	46.5% (53/114)	46.5% (53/114)
その他の	51.0% (25/49)	51.0% (25/49)	— (0/0)	0.0% (0/49)
から	93.8% (30/32)	93.8% (30/32)	— (0/0)	0.0% (0/32)
並びに	34.6% (9/26)	36.0% (9/25)	23.1% (6/26)	24.0% (6/25)
（読点のみ）	— (0/0)	0.0% (0/22)	— (0/0)	0.0% (0/22)
その他	38.5% (5/13)	38.5% (5/13)	1.7% (1/59)	7.7% (1/13)
かつ	20.0% (2/10)	15.4% (2/13)	20.0% (1/5)	7.7% (1/13)
と	— (0/0)	0.0% (0/5)	0.0% (0/163)	0.0% (0/5)

6.3.1　並列キーの種類別の解析結果

提案手法は，法令文の記述規則 [6][7] に適合した並列構造解析を実現するために，対象とする並列キーを従来手法のものから変更し，一部の並列キーに対して並列句候補の検出に関する新しい制約を加えた．具体的な改良点は次の (a) から (d) である．

(a)　「その他」と「その他の」を異なる並列キーとして扱う（5.1 節）．
(b)　「や」と「と」を解析対象の並列キーから外す（5.1 節）．
(c)　「から」を解析対象の並列キーに加える（5.1 節）．
(d)　「又は」,「及び」による並列構造の解析において並列句候補を検出するための制約を追加する（5.2 節）．

これらの改良は，並列キーの種類ごとに影響を与える．そこで，提案手法と従来手法をより正確に比較評価するため，並列キーの種類別の精度・再現率を求めた．表 2 にその結果を示す．提案手法は，解析対象としたすべての並列キーにおいて，精度・再現率ともに従来手法を上回っており，法令文の並列構造解析における本手法の有効性を確認できる．

以下では，上述した四つの改良点についてそれぞれ着目し，その効果について考察する．

(a)「その他」と「その他の」の区別による効果

従来手法において，「その他の」による並列構造の並列キーはすべて「その他」として検出される．そのため，正解データ中の「その他の」による並列構造 49 個をすべて誤って同定した．提案手法は正解データ中の「その他の」による並列構造の並列キーをすべて正しく検出し，同定した並列構造の精度・再現率はともに 51.0%を達成した．以

Vol.2017-NL-231 No.19
Vol.2017-SLP-116 No.19
2017/5/16

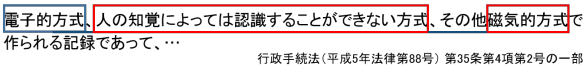

図 15 「その他」に先行する並列句と「その他」に後続する並列句を入れ替えた文

上より，「その他の」を区別した効果を確認できた．

しかし，提案手法においても，「その他の」や「その他」による並列構造の精度・再現率は全体の平均（それぞれ 66.1%, 62.6%）と比べて低かった．「その他の」の精度・再現率が低い理由として，「その他の」による並列構造は特殊な並列構造であるため流暢性スコアを計算できないこと，また，並列句の一つが必ず「その他」となるため類似性スコアが適切に算出できなかった可能性があることが考えられる．

また，「その他」の精度・再現率が低い理由として，「その他」に後続する並列句が，「その他」に先行する並列句と並列関係を持つだけでなく，先行する並列句の上位概念も表していることが考えられる．すなわち，これらの語句を入れ替えたことにより意味的に出現しにくい文となり，流暢性スコアを導入した意図に反し，このスコアが低く算出された可能性がある．例えば，図 3 に現れる並列句「磁気的方式」と「人の知覚によっては認識することができない方式」を入れ替えると，図 15 に示す文となる．したがって，「その他」による並列構造は「その他の」による並列構造と同じく，流暢性スコアを計算する際に何らかの配慮を払う必要があると考えられる．

(b) 「と」と「や」の除外による効果

正解データにおいて，「や」による並列構造は存在しなかった．また，「と」による並列構造は 5 個存在したが，従来手法はそのすべてを正しく同定できなかった．これらのことから，「と」や「や」を対象から除外することの悪影響はなかったと考えられる．なお，本実験で用いた従来手法の実装では，格助詞の「と」も並列キーとして抽出したため，精度を大きく低下させる要因となった．

(c) 「から」の追加による効果

正解データにおいて，「から」による並列構造の数は 32 個存在し，「並びに」，「その他」，「かつ」による並列構造の数よりも多い．また，「から」による並列構造に対しては，高い精度・再現率（ともに 93.8%）を達成した．そのため，「から」による並列構造を解析対象に加えることは有効であったといえる．

(d) 「又は」，「及び」に対する追加制約の効果

提案手法は，並列キー「又は」，「及び」のどちらにおいても，従来手法と比べて高い精度・再現率を達成した．この結果は，5.2 節の追加制約による効果ともいえるが，5.3 節で NLM を導入した効果も含まれている．

そこで，並列キー「又は」と「及びに」に対する制約を追加した効果を評価するため，これらの追加制約を提案手法から取り除いた手法により解析を行った．表 3 に，追加制約有り（提案手法）と追加制約無しのそれぞれに対して，並列キーごとの並列構造単位での精度・再現率を示す．「又は」による並列構造は 6 個，「若しくは」による並列構造は 4 個を新たに正しく解析できており，「又は」，「及び」に対する追加制約は，ある程度の効果があったといえる．

表 3 追加制約の効果

並列キー	追加制約有り（提案手法）		追加制約無し	
	精度	再現率	精度	再現率
又は	65.0% (165/254)	62.7% (165/263)	61.2% (159/260)	60.5% (159/263)
及び	77.2% (88/114)	77.2% (88/114)	73.7% (84/114)	73.7% (84/114)

表 4 階層別の評価結果

高さ	提案手法		従来手法	
	精度	再現率	精度	再現率
1	74.5% (369/495)	72.1% (369/512)	42.1% (256/608)	50.0% (256/512)
2	48.8% (63/129)	41.4% (63/152)	15.7% (22/140)	14.5% (22/152)
3	27.9% (12/43)	30.8% (12/39)	19.4% (7/36)	17.9% (7/39)
4	18.2% (2/11)	18.2% (2/11)	0.0% (0/8)	0.0% (0/11)
5	0.0% (0/1)	0.0% (0/3)	0.0% (0/1)	0.0% (0/3)

6.3.2 並列構造の階層別の解析結果

階層的並列構造は法令文の特徴の一つであるので，法令文の並列構造解析ではこれらを正しく解析できることが重要である．そこで，並列構造が属する階層別に評価を行う．

表 4 に並列構造が属する階層別で集計した並列構造単位の精度・再現率を示す．ただし，ここでは，正解データと並列句の範囲が一致するだけでなく，階層も一致した場合に正しく解析できたと判定する．すべての階層を通して，提案手法は従来手法よりも高い精度・再現率を達成しており，法令文に特有な階層的並列構造の解析において，本手法の有効性を確認できる．

しかし，提案手法の精度・再現率は，階層が高くなるにつれて低下した．その理由として，次のことが考えられる．

- 下位の並列構造の同定に失敗した場合，その影響を受ける．
- 階層が高いほど，その並列構造を形成する並列句は下位の並列構造を含むことになるため，単語数が長くなる傾向がある．そのため，並列句の候補が増加することになり，誤った候補を選択する可能性が高くなる．
- 図 14 で示した解析失敗例のように，アルゴリズム上正しく同定できない階層的並列構造が存在する．

したがって，階層が高くなるほど解析が難しくなるといえる．

6.3.3 並列句の数別の解析結果

すべての並列句の範囲を同定する必要があるため，一般に，並列句の数が多い並列構造であるほど，並列構造全体の解析は失敗する可能性が高まると考えられる．特に，提案手法と従来手法はともに，最右の並列句より一つずつ決定的に同定していくため，ある並列句の同定に一度でも失敗すると，その失敗を回復できず，それより左側にある並列句を同定することはできない．そのため，表 1 に示したように，提案手法の同定性能が従来手法を大幅に（F 値にして約 25%）上回っていたとしても，多くの並列句より構成される並列構造に対する同定性能が改善したかどうかは明らかではない．

そこで，並列句を多く持つ並列構造に対する本手法の有効性を検証するため，並列構造が持つ並列句の数ごとに，提案手法と従来手法の解析結果を比較評価した．表 5 に並列句の数ごとの精度・再現率を示す．精度・再現率ともに，並列句の数がどの場合（並列句の数が 7 のときの再現率を除く）においても，提案手法は従来手法を大幅に上回った．図 16 に 9 個の並列句からなる並列構造に対する同定結果を示す．この並列構造が持つ並列句はすべて法律名であり，そのほとんどが 1 文節で構成されているが，「民事訴訟費用に関する法律」のみ 3 文節から構成されている．従来手法はこの並列句の同定に失敗したため，残りの並列句を同定できなかった．一方，提案手法はこの並列句を正しく同定し，残りの並列句も正しく同定できた．以上より，並列句を多く持つ並列構造に対する本手法の有効性を確認した．

ところで，提案手法は並列句の数が増加するほど精度・再現率が向上する傾向があった．この傾向は，並列句の数が多いほど並列構造全体の解析は難しくなるという予想に反する．この原因を調査したところ，実験データにおいて，多数の並列句より構成される並列構造のほとんどは，図 16 のように，1 文節からなる並列句が多数を占め，複数文節からなる並列句は 1〜2 個含まれるという構成になっていることが分かった．具体的には，5 個以上の並列句から構成される並列構造 39 個のうち，32 個もの並列構造がこの構成に当てはまる．1 文節からなる句の並列関係の同定は比較的容易であることに加え，6.2 節の図 13 で示したように，提案手法は，文節数が異なる並列句が含まれていても，NLM によるスコアリングによって対処することができたため，これらの並列構造の多くを正しく同定することができたと考えられる．

一方，従来手法は並列句の数が増加するほど精度・再現率が低下する傾向があった．6.2 節の図 13 で示したように，従来手法は，単語数の異なる並列句の同定を苦手としているため，そのような並列句が途中に含まれていると並列構造全体の解析に失敗する傾向にある．上述したように，多数の並列句より構成されている並列構造は，1 文節からなる並列句のほかに，複数文節からなる並列句も少数含んでいるため，従来手法は失敗することが多くなったと考えられる．

表 5　並列句の数別の評価結果

並列句の数	提案手法		従来手法	
	精度	再現率	精度	再現率
2	64.2% (357/556)	60.4% (357/591)	37.0% (253/684)	42.8% (253/591)
3	63.2% (43/68)	65.2% (43/66)	31.3% (21/67)	31.8% (21/66)
4	75.0% (15/20)	71.4% (15/21)	27.8% (10/36)	47.6% (10/21)
5	90.0% (9/10)	81.8% (9/11)	25.0% (1/4)	9.1% (1/11)
6	100% (22/22)	91.7% (22/24)	— (0/0)	0.0% (0/24)
7	100% (1/1)	50.0% (1/2)	50.0% (1/2)	50.0% (1/2)
8	100% (1/1)	100% (1/1)	— (0/0)	0.0% (0/1)
9	100% (1/1)	100% (1/1)	— (0/0)	0.0% (0/1)

7. まとめ

本稿では，NLM を用いた法令文の並列構造解析手法を提案した．法令文を対象とした実験の結果，従来手法と比べてより正確に解析できることを確認した．また，結果に対する詳細な分析によって，従来手法からの改良がおおむね効果的に作用していることが明らかとなった．今後は，階層的並列構造の同定順序に関する問題（図 14）に対処し，言語モデルやスコアリング関数の精緻化を行うことで，更なる性能向上を図る．

参考文献

[1] 山田大介，島津明：法令文の言語的特徴を利用した可読性向上のための表示，言語処理学会第 12 回年次大会発表論文集，pp.196–199（2006）．
[2] 萩原正人，小川泰弘，外山勝彦：グラフカーネルを用いた非分かち書き文からの漸次的語彙知識獲得，人工知能学会論文誌，Vol.26，No.3，pp.440–450（2011）．
[3] 河原大輔，黒橋禎夫：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，言語処理学会第 13 回年次大会発表論文集，pp.506–509（2007）．
[4] 黒橋禎夫，長尾眞：長い日本語文における並列構造の推定，情報処理学会論文誌，Vol.33，No.8，pp.1022–1031（1992）．
[5] 松山宏樹，白井清昭，島津明：法令文書を対象にした並列構造解析，言語処理学会第 18 回年次大会発表論文集，pp.975–978（2012）．
[6] 石毛正純：自治立法実務のための法制執務詳解〔四訂版〕，pp.538–598，ぎょうせい（2004）．
[7] 大島稔彦：第 3 次改訂版　法制執務の基礎知識，pp.262–307，第一法規（2005）．
[8] Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C.:

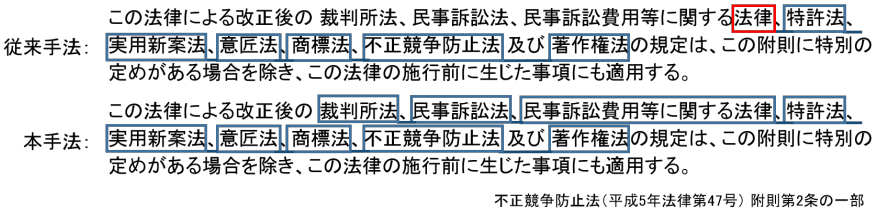

図 16 多数の並列句よりなる並列構造の解析例

A Neural Probabilistic Language Model, *JMLR*, Vol.3, pp.1137–1155 (2003).

[9] 山田将之，小川泰弘，外山勝彦：構文情報付き法律文コーパスの設計と構築，第 14 回言語処理学会年次大会発表論文集，pp.605–607（2008）.

[10] Sundermeyer, M., Schleuter, R. and Ney, H.: LSTM Neural Networks for Language Modeling, Proc. *INTERSPEECH 2010*, pp.194–197 (2010).

[11] Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, Proc. *EMNLP 2004*, pp.230–237 (2004).