バイパス付きアラインメントグラフを用いた日本語並列句検出と範囲同定

Bypassed Alignment Graphs for Detection and Scope Disambiguation of Japanese Coordinate Phrases

大熊 秀治
Hideharu Okuma

奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
okuma.hideharu01@is.naist.jp

原 一夫
Kazuo Hara

(同 上)
kazuo-h@is.naist.jp

新保 仁
Masashi Shimbo

(同 上)
shimbo@is.naist.jp

松本 裕治
Yuji Matsumoto

(同 上)
matsu@is.naist.jp

keywords: natural language processing, coordination, sequence alignment

Summary

We propose a machine learning-based method for analyzing coordinate structure in Japanese sentences. Effective methods for disambiguating coordination scopes already exist for English, but these methods assume input sentences always contain coordinations. Since detecting coordinations is non-trivial in Japanese, this assumption is often violated. The proposed method mitigates this problem by detecting the presence of coordinations and disambiguating their scopes simultaneously. It is built upon the previous work on English coordination that uses alignment graphs to evaluate the similarity of conjuncts. A “bypass” is introduced in these graphs to explicitly represent the non-existence of coordinations in a sentence, so that the feature weights for coordinations are learned separately from the weights for sentences not containing coordinations. We also propose to make all features dependent on the distance between conjuncts. In an experiment with the EDR corpus, the proposed method outperforms existing methods.

1. は じ め に

1.1 背 景

日本語の構文解析，特に係り受け解析は，情報抽出や機械翻訳といったアプリケーションの基盤技術として認識され，盛んに研究されてきた．初期の研究では，文節間の係りやすさを記述する規則を手手で整備していたが，構文情報の付いた大規模コーパスが入手可能になって以降，機械学習に基づく統計的な手法が主流になりつつある．係り受け解析においては，機械学習によって規則ベースの手法を上回る実用的な精度が得られている一方，未解決のタスクも残っており，その一つに並列句の曖昧性解消がある．並列句は高頻度で出現するため，その曖昧性を正しく解消することは優先度の高い課題である．

並列句は「と」や「」（読点）などの手がかり表現の前後に，名詞句や動詞句などが並列する構造であり，日本語の場合そこには 2 種類の曖昧性がある．一つは範囲の曖昧性であり，もう一つは手がかり表現の曖昧性である．

まず，並列句範囲の曖昧性について説明する．助詞「と」

を含む同一文に対する二つの異なる解釈を (1) に示す．文中の下線は，並列句を構成する単語列 (conjunct; 以下これを並列項目と呼ぶ) の範囲を表す．

(1a) 温暖化の抑制と経済の成長

(1b) 温暖化の抑制と経済の成長

このうち，(1a) が正しい解釈である．誤っている (1b) の解釈に従うと「温暖化」が「抑制」と「経済」の両方を修飾することになるが「温暖化の経済」は意味が通らない．

一方の手がかり表現の曖昧性は，並列句の存在を示唆するとされる表現が，常に並列句を導くとは限らないことを指す．たとえば，以下の二つの文

(2) 二条城と清水寺へ行った

(3) 友達と清水寺へ行った

は，どちらも並列句の手がかりとなる助詞「と」を含んでいるが，文 (2) の「と」は「清水寺」と「二条城」が並列関係にあることを示す (接続助詞) のに対し，(3) の「と」は並列関係を示さない (格助詞) ．

日本語並列句解析の代表的研究である、黒橋と長尾による規則ベースの手法 [黒橋 92] では、並列句を検出するための手がかり表現と並列句範囲を同定するための規則を手で定義している。この手法は、係り受け解析器 KNP [黒橋 00b] に組み込まれ、以来、新聞記事コーパスの整備と並行して、規則の拡張、重みの調整が随時行われてきた [黒橋 00a]。規則ベースの手法で実用的な解析性能を達成するためには、膨大な時間と労力が必要であることがうかがえる。他の言語、あるいは特許・法令文など異なる分野の文書への適用には、規則の再整備のためにさらに大きな労力が必要と予想される。

英語並列句解析では、原ら [原 07] が機械学習を利用する手法を提案している。この手法は系列アラインメントに基づく多数の素性を使用して、並列項目どうしの類似度を捉え、各素性の重みを訓練データに基づいて自動調整する。しかしながら英語には、手がかり表現の曖昧性がほとんどない（すなわち、“and”、“or” などの手がかり表現の周辺には、必ず並列句が存在する）ため、原らの手法が日本語においても十分な精度を示す保証はない。実際、5 章の実験において彼らの手法を日本語に素朴に適用した場合、解析精度は不十分なものであった。

この解決法として、文内の並列句の検出（手がかり語の曖昧性解消）を行った後、検出された文のみを対象に範囲の曖昧性解消を行う方法が考えられるが、このような逐次処理は現実的ではない。例 (2), (3) で見たように、並列項目間の類似度は、範囲同定のみならず検出の際にも重要な手がかりとなる。けれども、逐次処理最初の並列句検出段階においては、まだ並列項目範囲が確定しておらず、どの範囲の句の類似度を測定すべきかが不明である。このため、そもそも並列句の検出と範囲同定は不可分で同時に行うのが自然であり、原らの手法を日本語に適用するためには手法自体の改良が必要となる。

なお、KNP においても、黒橋・長尾の並列解析処理が出力した並列項目範囲のスコアが十分大きくない場合には「並列なし」と見なす [黒橋 00b] ことから、検出と範囲同定を同時に行っていると解釈できる。

1.2 本論文の貢献および構成

本論文では、原らの並列句解析法を日本語に適用するための改良法を提案する。具体的には、系列アラインメントを表現するためのアラインメントグラフ（編集グラフ）[Gusfield 97] に、グラフの始点と終点を結ぶバイパスと呼ぶ辺を新たに追加し、並列句の検出に利用する。すなわち、並列句が存在しない／存在するの区別を、グラフ内で最大スコアを持つ経路がバイパスを通る／通らないによって表現する。これによって、並列句の検出と範囲同定の両方を同時にかつ効果的に行うことができる。

また、原らは英語並列句のなかでも特に名詞-名詞並列だけを範囲同定の対象としているが、本研究では、より現実的な設定のため、あらゆる種類の（日本語）並列句

（例えば、名詞-名詞並列、動詞-動詞並列、動詞-形容詞並列）を扱う。名詞-名詞並列は短い文節数で構成されることが多いため、その範囲同定は比較的容易であるが、対象を名詞-名詞並列に限定しない本研究では、やや難しい問題を扱うことになる。この対処として我々は、並列句の検出および範囲同定に役に立つ素性は、並列句を構成する文節列の長さに応じて異なるという仮説を立て、その長短によってすべての素性を細分化する工夫を行う。

本論文は以下の構成をとる。2 章で関連研究を紹介し、日本語並列句解析の先行研究である黒橋と長尾の手法と、提案手法の基礎となる原らの英語並列句解析法について述べる。3 章で本研究が提案する日本語並列句の解析手法について説明し、用いる素性の概要を 4 章に記す。5 章で EDR コーパスを用いた評価実験の結果を示し、6 章で考察する。最後に 7 章で本論文をまとめる。

2. 関連研究

並列句範囲の曖昧性解消においては、並列項目の文法構造・意味的な類似度が大きな手がかりとなる。過去の研究では、この類似度を的確に捉える努力がなされてきた。

多くの研究では、解析対象を特定の形の句に限定し、その内部に生じる並列項目範囲の曖昧性解消を行っている。[Resnik 99] は「名詞₁ and 名詞₂ 名詞₃」という形の句における二つの解釈「(名詞₁) and (名詞₂ 名詞₃)」および「(名詞₁) and (名詞₂) 名詞₃」に関する曖昧性解消を行い、名詞₁ と 名詞₂ あるいは 名詞₁ と 名詞₃ の単複の一致や、コーパスから計算するそれらの意味的類似度が有効と報告している。[Nakov 05] はウェブ検索による単語の共起頻度を、[Chantree 05] はコーパス上の文脈（分布）類似度をそれぞれ利用し、同様の問題に取り組んでいる。機械学習を用いた研究としては、この問題を系列タグ付けタスクとみなし、条件付き確率場 (conditional random fields) を適用した例がある [Buyko 07]。

少数ではあるが本研究と同様、解析対象を短い句に限定せず、文全体から並列句範囲を同定する研究もある。これらのうち本研究が直接関係する二つの先行研究については、2.1 節、2.2 節に詳しく述べる。この他に日本語並列句解析に関する最近の論文として、大規模コーパスを用いた [Kawahara 07, Kawahara 08] が挙げられる。

2.1 黒橋と長尾の規則ベース日本語並列句解析法

黒橋と長尾 [黒橋 92] による、規則ベースの並列句解析法（以下黒橋・長尾法）は、日本語係り受け解析器 KNP に内部処理として組み込まれ [黒橋 00b]、広く使われている。KNP ではまず、人手で収集された並列手がかり表現（並列のキー）が入力文に含まれるか調べ、含まれていた場合に、黒橋・長尾法が起動される。黒橋・長尾法は、特定したキー文節の前後のあらゆる文節列対に対し、そのうちのどれが並列する二つの文節列であるかを決定

図1 “a policeman and warehouse guard” に対するアラインメントグラフ (a) と、二つの異なる並列句範囲を表す経路 (b, c), および、並列句なしとする解釈に相当する経路 (d)。

するためのスコアを計算し、スコア最大となる文節列の対を、並列項目の範囲として出力する。ただし、スコアが一定の基準値に達しない場合には、並列句は存在しない、と判定し終了する。スコアは、あらかじめ人手によって決定された以下の3種類の得点の和として定義される。

- 文節間の類似度：二つの文節列において、並列対応する文節間の類似度を測る基準（例：自立語の品詞が一致すること）に対して与える加算点
 - 区切り要素の有無：並列句をなす文節列の中にあまり現れない区切り要素（例：「～は」）に対する減点
 - 範囲マーカの有無：並列句の範囲の内側と外側を分けるマーカ（例「など」）に対するボーナス点
- 黒橋と長尾は、このようなスコア最大となる並列項目の探索が動的計画法を用いて効率良く行えることを実証した。

2.2 原らの英語並列句解析

原らの並列解析法 [原 07] の基礎となるのは、遺伝子等の系列アラインメントに用いられるアラインメントグラフ (alignment graph; あるいは編集グラフ edit graph) [Gusfield 97] と呼ばれる格子状の有向グラフである。このグラフには、水平、垂直、対角方向の3種類の辺が存在し、各々、右、下、右下方向にのみ遷移が許される。対角方向の辺は、その行方向の文字を列方向の文字に置換する編集操作を表す。水平方向の辺は列方向の文字を挿入する編集操作に、垂直方向の辺は行方向の文字を削除する編集操作に、それぞれ対応する。

ただし、並列句解析では1本の系列(文)内の類似度が高い部分系列(すなわち並列句)を探索するのが目的であるため、グラフは矩形ではなく、上三角形^{*1}をし

*1 黒橋と長尾が用いた動的計画法もまた三角形であるが、並列

ている。また、文字ではなく単語単位でのアラインメントを考えるため、格子の行および列には解析対象の単語列(すなわち文)を配置する。なお、本論文で扱う日本語文の場合、単語ではなく文節を割り当てる。例として“a policeman and warehouse guard”という単語列に対するアラインメントグラフを図1(a)に示した。行方向の単語は各辺の右に、列方向の単語は各辺の上に表示してある。

左上隅と右下隅の節点を、それぞれグラフの始点と終点と呼ぶ。文内の並列構造は始点から終点に至る経路として表現される。経路上の辺は、編集操作を受ける語が並列句範囲の内側にあるか外側にあるかに応じて、InsideあるいはOutsideのラベルをもつ。例として、図1(a)グラフ上の三つの経路を図1(b)–(d)に示す。各々の経路は“a policeman and warehouse guard”の並列句範囲に関する異なった解釈を表している。図中、Insideラベルの辺(以下単にInside辺と呼ぶ)を実線で、Outsideラベルの辺(Outside辺)を点線で描いた。並列句の項目が“policeman”と“warehouse guard”であることを示すのは図1(b)の経路であり、“policeman”と“warehouse”とするのは図1(c)の経路である。並列句が存在しないとする解釈は、図1(d)の経路となる。Outsideラベルは水平方向と垂直方向の辺にのみ許されること、垂直方向の辺から水平方向の辺への接続は同じラベルの場合は禁止されている [原 07] ことから、Outsideラベルの辺のみからなる経路は、グラフ右上節点を通る経路に一意に定まる。

原らは、アラインメントグラフの辺および接続する2辺に対して多数の素性を割り当てたうえで、グラフ中の経路のスコアを、経路上の辺に出現した素性の線形関数と定義し、各素性の重みを訓練データを用いて自動調整した。 x を解析対象の文、 e を x に対するアラインメントグラフ上の辺または接続する2辺を表す変数^{*2}とし、 f_i ($i = 1, \dots, M$; M は全素性の数) を素性関数とする。ここで各 f_i は、 x と e の関数であり、 e に割り当てられた素性の値は $f_i(x, e)$ と書ける。素性 f_i が辺(または接続する2辺) e に割り当てられていなければこの値は0とする。 i 番目の素性 f_i の重みを w_i と書き、 π をアラインメントグラフ上の経路とすると、 π のスコアは

$$\text{score}(\pi) = \sum_{e \in \pi} \sum_{i=1}^M w_i f_i(x, e)$$

で与えられる。ただし $e \in \pi$ は e が経路 π に含まれることを表す。各素性重み w_i は文によらずシステム全体で共通であり、各訓練データ中に明記された正しい並列句に対応する経路のスコアが、グラフ上の他のすべての経路のスコアよりも大きくなるよう、平均化パーセプトロン (averaged perceptron) アルゴリズム [Collins 02] を用

キーを固定した後でその右上の矩形領域に探索範囲を限定して探索を行う点で、2種類のラベルで並列句内外を区別し、全体経路を探索する原らの手法と異なっている。

*2 e は具体的には、辺(の接続)のグラフ内の位置、方向(水平、垂直、対角)、ラベル(Inside, Outside)の情報を表す。

図2 複数の並列句を含む文に対するアラインメントグラフ

図3 バイパス付きアラインメントグラフ. 鎖線がバイパス辺.

いて調節される．なお，訓練データ作成コストの削減などの目的で，単語／文節単位のアラインメント（アラインメントグラフ上の経路）ではなく，並列句の範囲を訓練データとして与えることがあり，本論文の実験もこの設定で行う．この設定への対処方法は，[原 07]に詳しい．

さて，黒橋・長尾法で用いられる規則の多くは，アラインメントグラフ上で等価な素性として表現できる．すなわち，文節間類似度に基づく得点は，対角方向の辺（Insideラベルを持つ）に割り当てる素性の重みに相当する．区切り要素に対する減点は，Inside 辺の接続に対する素性の重みと関係がある．並列句の検出規則および範囲マーカーに対するボーナス点は，Outside 辺と Inside 辺の接続，あるいは Inside 辺と Outside 辺の接続に対する素性の重みに関連する．さらに，「各文節対の種類に応じた個別の類似度得点」のような，数が膨大なため規則ベースの手法で用いるのが困難なものについても，機械学習を用いた手法では，それらすべてを素性としてモデルに組み込み，重みを自動調整することができる．並列句検出のための並列のキーも，事前に定義する必要はなく，訓練データから学習できる．

3. 提案手法

本章では，原らの英語並列句解析法を改良した，日本語並列句解析の手法を提案する．まず，1章で述べた，日本語並列句の検出自体が容易でないという問題に対処するために，原らのアラインメントグラフに対して変更を加える．さらに素性について，並列句をなす文節列の長短による細分化を行う．各々について以下で説明する．

3.1 バイパス付きアラインメントグラフ

原らの手法では，解析対象文に並列句が含まれないとする解釈を，図 1(d) で見たように，Outside 辺のみから

なる経路として表現する．ここでは，経路上の Outside 辺は，解析対象文に並列句が含まれるか否かを判定する役割（並列句の検出の役割）を担う．すなわち，Outside 辺のみからなる経路のスコア（＝個々の Outside 辺に割り当てられている素性重みの総和）が，Inside 辺を含む経路のスコアよりも大きいとき，解析対象文に並列句は含まれないと判定する．

他方，Outside 辺は，グラフの（左上）始点と（0 個以上の）並列句範囲を示す Inside 辺と，グラフ（右下）終点を繋ぐための便宜上の辺としても用いられるため，図 1(b)(c) や，図 2 に例示する複数の並列句を表す経路に見るように，並列句の存在を表す経路内においても必ず出現する．この場合の Outside 辺は，編集操作を受ける語（垂直方向の Outside 辺なら右，水平方向の Outside 辺なら上に位置する語）が，並列項目ではないことを示す役割（並列句の範囲同定の役割）を担っている．

このように，原らは Outside 辺に対し，並列句の検出と並列句の範囲同定という，目的の異なる複数の役割を与えている．我々は，この混用が原らの手法が日本語並列句解析において十分な精度が得られない一因と考える．そして，この問題の解消のため，並列句が含まれていない文を表す専用の経路をアラインメントグラフに導入する．すなわち，グラフの始点と終点を直接つなぐバイパスと呼ぶ辺を追加し（図 3），この辺を通る経路をバイパス経路と呼ぶ．さらに，アラインメントグラフ右上隅の節点において，左から入る Outside 辺と下に向かう Outside 辺との接続を禁止することで，図 1(d) で示されたような Outside 辺のみからなる経路がグラフ内に存在できないようにする．この変更により，Outside 辺は並列句の存在を表す（Inside 辺が少なくとも一つ含まれる）経路を連結する役割のみを担い，バイパス経路によって表される非並列文の検出とは役割が分離される．

結局，我々が提案するバイパス付きアラインメントグラフを用いると，バイパス経路がそれ以外のすべての経路（＝Inside 辺を必ず含む）より高いスコアを持つ場合に，「解析対象文に並列句が含まれない」と判定される．

バイパス経路の素性としては，文の長さや，文に出現する助詞といった，文全体の特徴を用いる．提案手法は，これらの素性の重みも単語（文節）単位の素性と全く同様に扱い，同時に訓練データを用いて自動調整する．

3.2 文節間の距離に応じた素性の細分化

提案手法が用いる素性は，すべて指標関数（indicator function; 条件に応じて 1 あるいは 0 を返す関数）として実装されている．これらの素性については 4 章で触れることとし，本節では，素性設計の新たな工夫，すなわち，文節間の距離に応じた素性の細分化について説明する．細分化の動機は，以下の観察による．

- 並列項目となる文節列は一般に短い．たとえば，EDR コーパス（平凡社百科事典セクション）では，約 85%

図 4 (a) 短い (文節数 3) 並列項目, および, (b) 長い (文節数 7) 並列項目からなる並列句のアラインメントグラフ上の経路表現. 経路上の Inside 辺 (実線) の, グラフ左端からの距離 d は並列句をなす文節列の長さに関連し, その長さが長いほど, 経路はグラフ左端から離れる.

が長さ 5 以下である.

- 並列項目が長い場合には, 類似度の高い文節の並用, 曖昧性の少ない並列のキーによる接続など, 並列句であることを示す強い特徴を伴うことが多い.

短い / 長い文節列対からなる並列句を含む文の例を, そのアラインメントグラフ上の経路とともに図 4 に示す. 図 4(b) の長い文節列の対は, 「不純物が」および「利用した」の共有, 4 文字一致する「コールドトラップ」と「ホットトラップ」の並用, さらに, 曖昧性の少ない並列手がかり語「あるいは」による接続など, 並列句の存在を示す強い特徴をもつ. 他方, 短い並列項目からなる並列句 (たとえば図 4(a)) においては, 必ずしもこのような顕著な特徴は見られない.

並列項目をなす文節列の長短によって並列句の特徴が異なるため, この長短に応じて素性を別々のものとして扱うことで精度の向上が期待できる. 並列項目の長さは, アラインメントグラフ上の経路が定まって初めて決まることから, その経路計算の基礎となる素性を並列項目の

長さに依存させるのは不可能である. しかしながら素性を, グラフ内の辺の位置, 具体的には, 辺の始点から, グラフの同じ行の左端までの距離 d に依存させることはできる. そしてこの距離は, 辺の水平方向にある文節と垂直方向にある文節の間の, 文内における距離 (二つの文節に挟まれた文節の数) に等しいため, 並列項目の長さに関連がある. たとえば, 図 4(a) に示す経路は並列項目長が 3 の並列句を表すが, この経路に含まれる Inside 辺の, グラフ左端からの距離 d はすべて 2 である. 一方, 並列項目の長さが 7 である図 4(b) では, Inside 辺のグラフ左端からの距離は 7 または 8 であり, 相関が見て取れる.

そこで, 本研究ではグラフ左端からの距離 d の大小により, 辺に割り当てる素性をすべて細分化する. すなわち, 条件 X が満たされたかを識別する既存の指標関数 f

$$f = \begin{cases} 1, & \text{if } X; \\ 0, & \text{otherwise.} \end{cases}$$

で定義される個々の素性に加えて, 新たに次の素性を導入する.

$$f' = \begin{cases} 1, & \text{if } X \text{ and } d \leq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

ここで θ は固定値として与える閾値である. このようにすべての素性を細分化することで, 素性の数は 2 倍に増えるが, 並列句の長短による特徴の違いを, それぞれ異なる素性の重みとして学習できるようになる^{*3}.

4. 日本語並列句解析のための素性

以下では, 5 章の実験で使用する素性について簡単に説明する. まず, 素性の定義に必要な文節の属性について触れた後, アラインメントグラフの辺 (あるいは接続する 2 辺) に割り当てる素性について述べる.

4.1 文節の属性

本論文では, 文節の主辞となる単語, 文節の「語形」(後述)となる単語, 文節内の助詞などの特徴 (表 1) を, 文節の属性と呼ぶ. 属性は個々の文節の特徴であるのに対し, 後述の素性は 1 個あるいは複数の文節の属性に基づいて定義され, グラフの辺 (あるいは接続する 2 辺) に対して割り当てられる, という違いがある.

主辞とは文節内で最も末尾近くに現れる自立語である. 語形^{*4}とは文節内で最も末尾に近い, 助詞および句読点以外の単語 (自立語とは限らない) を指す. 主辞と別に語形に関する属性を導入する理由は, 日本語においては接尾語が活用形を持つことがあり, 並列句解析の手が

^{*3} 別の方法として, f' の導入だけではなくさらに「 X かつ $d > \theta$ 」の時だけ値 1 をとる別の指標関数 f'' で f を置き換えてしまうことも考えられる. 予備実験では, この場合でも提案した f, f' を用いる手法と大差ない精度が得られている.

^{*4} 「語形」という呼称については文献 [工藤 02] に倣った.

表 1 素性定義に使用する文節の属性

名前	内容
BaseForm	主辞の基本形
Pos	主辞の品詞
PosSubcat	主辞の品詞細分類
ConjForm	主辞の活用形
Type	主辞が体言か用言かそれ以外か
BunruiCode	主辞のシソーラスコード
AuxBaseForm	語形の基本形
AuxPos	語形の品詞
AuxPosSubcat	語形の品詞細分類
AuxConjForm	語形の活用形
Particle	文節内の助詞の表層と品詞細分類
Punctuation	文節内の句読点の有無と種類
Digit	文節内に数値表現を含むか
ProperNoun	文節内に固有名詞を含むか
Location	文節内に場所名を含むか
Person	文節内に人名を含むか
Organization	文節内に組織名を含むか
Surface	文節の表層文字列

りとして主辞ではなく語形の活用形を用いた方が、場合によっては適切なためである。なお、今回は比較実験の都合上、品詞や活用形、場所・人名・組織名等の情報は形態素解析器 JUMAN の出力に基づいている。

4.2 素性の例

素性はいずれも、特定の条件が満たされるか否かに関する指標関数として実現される。例えば次の文

(4) 広い邸宅を持ち, 多くの使用人を雇っている。

の二つの並列項目（下線部）は動詞（「持ち」「雇っている」）を含むという類似性を持つ。この類似性は、アラインメントグラフ中の各辺 e に対して、

$$f(\mathbf{x}, e) = \begin{cases} 1, & \text{if } (e \text{ のラベル}) = \text{Inside} \\ & \text{and } (e \text{ の方向}) = \text{対角方向} \\ & \text{and } \mathbf{x}[i].\text{Pos} = \text{動詞} \\ & \text{and } \mathbf{x}[j].\text{Pos} = \text{動詞}; \\ 0, & \text{otherwise.} \end{cases}$$

という指標関数を素性として導入し捉えることができる。ここで、 (i, j) は e の始点のグラフ内座標（すなわち e の始点は、グラフの、上から i 行目、左端から $(j-1)$ 列目（文中 j 番目文節に対応する）の桁の左上節点）とする。また、 $\mathbf{x}[i].\text{Pos}$ は入力文 $\mathbf{x}$ における i 番目の文節の Pos 属性の値を表す。ただし実際には、上で掲げた特定の属性と属性値のみではなく、各属性 a と、その取りうる値 u, v の組み合わせ全てに対して、

$$f_{a,u,v}(\mathbf{x}, e) = \begin{cases} 1, & \text{if } (e \text{ のラベル}) = \text{Inside} \\ & \text{and } (e \text{ の方向}) = \text{対角方向} \\ & \text{and } \mathbf{x}[i].a = u \text{ and } \mathbf{x}[j].a = v; \\ 0, & \text{otherwise.} \end{cases}$$

という素性を導入している。素性の総数は結果として膨大になるが、各素性の重みは訓練データに基づいて機械学習（平均化パーセプトロン）によって調整できる。

スペースの制約上、具体的な素性の一覧は、テクニカルレポート [Okuma 09] に譲る。以下では用いた素性のうち、外部リソース（シソーラス、コーパス上の共起頻度）に依存する二つの素性、および、バイパス辺に割り当てる素性について解説する。

§1 シソーラス素性

並列項目間の類似度を測る補助としてシソーラスに基づく素性を導入する。シソーラスには、[黒橋 92] にならない分類語彙表 [国立 04] を用いた。分類語彙表では各語に 5 桁の分類番号が与えられており、各桁の値はシソーラス木における語の位置（語が所属する意味的なクラス）を表している。我々はこの分類番号に基づく 2 種類の素性を用いる。一つ目は、二つの文節主辞がシソーラス木のどの深さまで同じ意味クラスに属しているかを表す素性である（特定の深さまで一致する場合に 1 を返す指標関数）。黒橋と長尾も同様に、シソーラス木において二つの主辞が一致する深さに応じて加点する規則を用いている。二つ目は、二つの主辞の分類番号の組み合わせ（の指標関数）である。シソーラス木の深さに依存する素性は、一致する深ささえ同じであれば同一の重みが与えられる。一方、分類番号の組み合わせに対する指標関数を用いることで、どの意味クラスの語の組み合わせが並列になりやすいのかを、各組み合わせごとに独立の重みとして学習することができる。

§2 共起素性

単語の共起頻度もまた並列句になりやすいか否かの手がかりとなる。「A の B と C が」という表現の解釈として、「A の B」、「A の C」の双方がコーパスに頻出するなら、並列句として「(A の B と)(C が)」という構造よりも「A の ((B と)(C が))」という構造を（他に手がかりがなければ）選ぶべきである。[Chantree 05, Kawahara 08] が同様の考察に基づいた特徴量を用いている。この手がかりをアラインメントグラフ中に素性として組み込むため、あらかじめ共起表現を毎日新聞記事 10 年分から収集した。記事を JUMAN で形態素解析し、係り受け解析器 KNP の“-bnst”オプションを用いて文節区切りのみ（係り受け解析は行わない）を行った後、隣接する文節中に出現する全ての自立語対を集め、この中から尤度比検定 [Dunning 93] によって抽出される自立語対を共起表現と定めた。素性としては、2 対の文節の主辞が、(上の基準に基づく) 共起表現か否かの指標関数を用いる。この素性はアラインメントグラフ上で、Inside 辺と Outside 辺の接続に対して割り当てる。

§3 バイパス辺に割り当てる素性

3.1 節で導入したバイパス経路に割り当てる素性は、文節単位ではなく文単位の素性となる。長い文ほど、そして、特定の助詞を含む文ほど並列句が現れやすいと考えられるため、文の長さに関する 5 個の指標関数（文の文節数 1-5, 6-10, 11-15, 16-20, 21 以上）、および文に出現する助詞の種類の指標関数を素性として用いる。

表 2 並列項目範囲と並列係り受け関係判定の結果. 括弧内の数値は, 原らの手法からの増減 (ポイント) .

手法	(i) 並列項目範囲			(ii) 並列係り受け		
	適合率	再現率	F 値	適合率	再現率	F 値
KNP	n/a	n/a	n/a	58.8	65.3	61.9 (-2.6)
原らの手法	62.4	58.0	60.1 (± 0.0)	67.0	62.1	64.5 (± 0.0)
原らの手法+素性細分化	63.7	60.1	61.8 (+1.7)	68.3	64.3	66.2 (+1.7)
原らの手法+素性細分化+バイパス	62.2	65.5	63.8 (+3.7)	66.8	69.9	68.3 (+3.8)

5. 評価実験

EDR 電子化辞書 1.5 に収録されたコーパス (EDR コーパス) の一部を用い, 提案手法を原らの手法と比較する. また, 黒橋と長尾の規則ベース並列句解析法が組み込まれている日本語係り受け解析器 KNP との比較も行う.

5.1 実験データ

EDR コーパスは複数のセクションからなるが, 本研究では並列句を含む文の割合が最も大きい平凡社百科辞典セクションを用いる. EDR コーパスでは, 文は単語区切りされ, 各単語に概念情報などが付与される. さらに, 各文には構文情報と意味情報が付与されている. 構文情報は S 式による依存構造の木として与えられ, 意味情報は構文木とは別に意味フレームとして与えられる. 並列句は, 意味フレームにおいて, 二つの単語を関係付ける概念関係子 “and” に対応する. この情報と構文木の情報を組み合わせることにより, 並列項目の範囲を特定できる.

平凡社百科辞典セクションの 10,072 文のうち, 5,880 文が並列句を含む. このうち, 入れ子の並列句を含む 1,791 文は, 提案手法, 原らの手法のどちらも取り扱えないため, 解析対象から除外する. 残る 8,281 文 (= 10,072 - 1,791) を形態素解析器 JUMAN 5.1^{*5} と係り受け解析器 KNP 2.0^{*6} により解析し, 文節区切りを行う. EDR コーパスに本来タグ付けされた形態素の情報や文節区切りを用いず, JUMAN, KNP で改めて解析する理由は, KNP (JUMAN の解析結果を入力として要求する) との比較を同一条件で行うためである. この文節区切りの結果, 並列句全体が一つの文節内に含まれてしまう場合と, EDR コーパスに明らかなタグ付け誤りがある場合, 文全体を解析対象から除外する. 以上の処理の結果として, 並列句を含む 3,257 文と, 並列句を含まない 4,192 文 (= 10,072 - 5,880) の, 合計 7,449 文を実験データとして得る.

5.2 比較対象の手法

原らの英語並列句解析法に対して日本語向けに必要最低限の変更を行ない, 4 章の素性を用いたものを以下「原らの手法」と呼び, 評価の基準とする. この手法ではバイパスの導入および距離に依存した素性の細分化は行われていないことに注意されたい. 評価基準である「原らの手法」に対して 3.2 節の「素性細分化」および 3.1 節

で提案したアラインメントグラフへの「バイパス」追加を行い, 精度の変化について調査する. なお, 素性細分化のための閾値 θ の値は 4 とした. 参考のため, あわせて KNP の精度について調査する.

5.3 評価基準

評価には次の二つの基準を用いる.

- (i) 並列項目範囲: 並列句の各並列項目範囲の両端を正しく同定できるか.
- (ii) 並列係り受け: 並列句内で連続する 2 並列項目の終端文節対を正しく同定できるか.

基準 (ii) は, KNP との比較のために用意した. KNP の出力する係り受け木において, 並列句は連続する並列項目対の終端文節間の依存関係 “P” によって表現される. この表現では並列項目の始点がわからず, 基準 (i) を評価に使うことができない. 一方, 並列句の終端については “P” 型依存関係の両端によって知ることができるため, 基準 (ii) による評価は可能となる. 評価尺度として, 適合率 (precision), 再現率 (recall) の調和平均である F 値 (F-measure) を用いるが, 傾向分析のため適合率, 再現率についても報告する.

5.4 実験結果

5 分割交差検定の結果を表 2 に示す. まず, 改良を加えない原らの手法が KNP を上回る精度を得ていることがわかる. KNP は並列句を積極的に検出する傾向があり, 並列句を実際には含まない文に対しても, 誤って並列句を検出しやすい. よって, KNP は高い再現率を得る一方で, その適合率は高くない. 他方, 原らの手法は適合率が高い反面, 再現率が低い.

原らの手法に対し, 文節列の距離に応じた素性細分化を行うと, 範囲同定の精度は F 値で 1.7 ポイント上昇する. 加えて, アラインメントグラフにバイパス経路を追加すると, F 値の上昇は 3.7 ポイントに達する. 原らの手法に対して再現率が顕著に改善されていることがわかる. 係り受け判定による評価では, F 値で KNP を 6.4 ポイント上回る.

6. 考察

本章では, 5.4 節の実験結果について分析し, 提案手法に必要な改良点について議論する.

*5 <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html>

*6 <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html>

表3 並列項目の長さ別の再現率

並列項目の長さ（文節数）	5 以下	6 以上
原らの手法	61.9	35.1
原らの手法+素性細分化	62.8	44.0

6.1 バイパス導入の効果と再現率

最初に、バイパスなしモデル（原らの手法，原ら+素性細分化）の再現率が低い理由について考察する．一般的に機械学習による識別法の出力は，訓練データのラベル分布に強く依存し，データ中で頻度の高いラベルを出力しやすい傾向がある．特に，素性設計が不適切で，素性から識別に有効な手がかりが得られない場合にこの傾向は顕著になる．本実験の訓練データ中の文は，その半数以上が並列句を含まないため，実際には並列句が含まれる文を誤って「並列句なし」と判定しやすく，再現率が低下する結果となったものと推測できる．このタイプの誤りは「原らの手法」で733文，「原ら+素性細分化」で697文ある．バイパス辺を追加した「原ら+素性細分化+バイパス」においては，同じ誤りが510文にまで減り，結果として（適合率を維持したまま）再現率が改善している．同一の素性（Outside 辺に割り当てられた素性）が担っていた並列句検出と範囲同定の二つの役割が，バイパス辺の追加によって分離され，より識別に適した素性設計となった結果と考える．

6.2 文節間距離に応じた素性の細分化の効果

素性の細分化により並列項目範囲同定の結果が改善した例を以下に掲げる．(5a) が原らの手法の出力する並列項目（下線にて表示），(5b) が原らの手法+素性細分化の出力する並列項目（正解）である．

(5a) この部分は光を正反射しなくなるので顕微鏡では暗く見え，比較的腐食液に侵されない部分は正反射して明るく見えるので，...

(5b) この部分は光を正反射しなくなるので顕微鏡では暗く見え，比較的腐食液に侵されない部分は正反射して明るく見えるので，...

この例^{*7}における正解並列項目は各6文節，7文節，と比較的長い．全データ中の正解並列項目を項目の長さで2分割（5文節以下 / 6文節以上）し，各分割ごとに評価した再現率を，表3に示す．並列項目の総数は，5文節以下が6336個（全体の85.5%），6文節以上が1076個（同14.5%）である．素性細分化によっていずれの再現率も向上しているが，特に長い並列項目に対する伸びが大きい．

^{*7} (5a) では並列項目が連続していないが，並列項目は文内で常に連続して現れるとは限らないため（「さらには」「たとえば」，といった挿入句が項目間に現れることがある），並列項目は連続しなければならないという制約は導入していない．このような制約を導入している[黒橋94]では，挿入句の取り扱い方法が課題としてあげられている．

図5 文(7)における並列項目範囲の経路表現．(a) 単一バイパス付きアラインメントグラフと (b) 複数バイパス付きグラフ

6.3 解析誤りの例と改善方法

「原ら+素性細分化+バイパス」の適合率，再現率は，項目範囲同定でそれぞれ62.2%/65.5%（表2参照）であり，英語並列句に対するCollinsパーザの精度62.2%/61.5% [Collins 99]と同程度であるが，さらなる精度向上は今後の課題である．以下で可能な改善方法を考察する．

§1 素性の改善

提案手法（原ら+素性細分化+バイパス）による代表的な解析誤り例を，以下に示す．(6a) 下線部が提案手法の出力並列項目^{*8}，(6b) 下線部が正解並列項目範囲である．

(6a) 温帯地域での森林減少やサバンナ地帯の乾燥化は，...

(6b) 温帯地域での森林減少やサバンナ地帯の乾燥化は，...

この誤りを解決するためには，[Kawahara 07, Kawahara 08]で用いられているような，より大規模なコーパス上での共起頻度を活用することが考えられる．つまり，「温帯地帯での」+「サバンナ地帯」，「森林減少の」+「乾燥化」がそれぞれ共起しにくいという情報を素性として与えられれば，上の誤りは避けられる可能性がある．

§2 複数のバイパスを持つアラインメントグラフ

提案したバイパス付きアラインメントグラフの問題として，以下のような状況を考慮していない点がある．文(3)（「友達と清水寺に行った」）は並列句を含まないが，これはバイパス付きアラインメントグラフ上でバイパス経路によって表現される．一方，以下の文

(7) 友達と清水寺に行った時に，ハツ橋と冷やし飴を買った

を考えると，「ハツ橋」「冷やし飴」が並列をなすため，グラフ上でバイパス経路は辿られない（図5(a)）．つまり，文(7)中の「友達と清水寺に行った時」と文(3)はほぼ

^{*8} 実際には文節単位の処理を行っているが，例文(6)では見やすさを優先し文節末の助詞は下線部から除外してある．

同一であるにもかかわらず，一方はバイパス経路が正解であり，他方はバイパス経路ではなく Outside 辺の接続が正解とされる．このため，上のような文の一方が訓練データ，他方がテストデータ中にあったとしても，正しい判断ができない可能性がある．

これに対しては，アラインメントグラフ中の（始点と終点だけではなく）全ての左端節点間にバイパス辺を設け，文 (7) に対しても文内の非並列部分についてはバイパス辺を通る経路として表現することが考えられる（図 5(b)）．この複数バイパス付きグラフの有効性の検証は今後の課題である．

7. む す び

原らの並列句解析法に対して，アラインメントグラフへのバイパスの追加と，素性の文節間距離による細分化という比較的簡単な修正によって，日本語における並列句解析精度が顕著に向上することを示した．

素性細分化は日本語に限らず，英語でも有効と考えられるため，検証を行う予定である．また，提案手法は独立した並列句解析法であり，特定の形態素解析器や構文解析器に依存しない柔軟性があるが，黒橋・長尾の並列解析法が KNP に組み込まれているように，係り受け解析との統合についても考えたい．

謝 辞

複数並列句を含む文に対する提案手法の欠点について，両査読者からの確かな指摘を受けた．謝意を表する．

◇ 参 考 文 献 ◇

- [Buyko 07] Buyko, K., Tomanek, K., and Hahn, U.: Resolution of coordination ellipses in biological named entities using conditional random fields, in *Proc. 10th PACLING*, pp. 163–171 (2007)
- [Chantree 05] Chantree, F., Kilgariff, A., de Roeck, A., and Willis, A.: Disambiguating coordinations using word distribution information, in *Proc. International Conference on Recent Advances in Natural Language Processing*, pp. 144–151 (2005)
- [Collins 99] Collins, M. J.: *Head-driven Statistical Models for Natural Language Parsing*, PhD thesis, Univ. Pennsylvania (1999)
- [Collins 02] Collins, M.: Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms, in *Proc. EMNLP 2002*, pp. 1–8 (2002)
- [Dunning 93] Dunning, T.: Accurate methods for the statistics of surprise and coincidence, *Computational Linguistics*, Vol. 19, pp. 61–74 (1993)
- [Gusfield 97] Gusfield, D.: *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press (1997)
- [原 07] 原 一夫, 新保 仁, 松本 裕治: アラインメントと機械学習を応用した並列句解析, 人工知能学会論文誌, Vol. 22, No. 3, pp. 248–255 (2007)
- [Kawahara 07] Kawahara, D. and Kurohashi, S.: Probabilistic coordination disambiguation in a fully-lexicalized Japanese parser, in *Proc. 2007 EMNLP-CoNLL Joint Conference*, pp. 306–314 (2007)
- [Kawahara 08] Kawahara, D. and Kurohashi, S.: Coordination disambiguation without any similarities, in *Proc. 22nd COLING*, pp. 425–432 (2008)
- [国立 04] 国立国語研究所: 分類語彙表, 大日本図書 (2004)
- [工藤 02] 工藤 拓, 松本 裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, pp. 1834–1842 (2002)
- [黒橋 92] 黒橋 禎夫, 長尾 真: 長い日本語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)
- [黒橋 94] 黒橋 禎夫, 長尾 真: 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)
- [黒橋 00a] 黒橋 禎夫: コーパスが先か, パーサーが先か, 情報処理学会, Vol. 41, No. 7, pp. 769–773 (2000)
- [黒橋 00b] 黒橋 禎夫: 結構やるな, KNP, 情報処理学会, Vol. 41, No. 11, pp. 1215–1220 (2000)
- [Nakov 05] Nakov, P. and Hearst, M.: Using the Web as an implicit training set: application to structural ambiguity resolution, in *Proc. HLT-EMNLP Joint Conference*, pp. 835–842 (2005)
- [Okuma 09] Okuma, H., Hara, K., Shimbo, M., and Matsumoto, Y.: Bypassed alignment graph for learning coordination in Japanese sentences: supplementary material, Technical Report NAIST-IS-TR2009003, Nara Institute of Science and Technology (2009)
- [Resnik 99] Resnik, P.: Semantic similarity in a taxonomy, *Journal of Artificial Intelligence Research*, Vol. 11, pp. 95–130 (1999)

〔担当委員: 河原 大輔〕

2009 年 7 月 22 日 受理

著 者 紹 介

大熊 秀治

2007 年 千葉大学工学部情報画像工学科卒業. 2009 年 奈良先端科学技術大学院大学情報科学研究科情報処理学専攻修了. 現在 (株) リコー勤務.

原 一夫

1994 年 東京大学工学部計数工学科卒業. 1996 年 同大学院工学系研究科修士課程計数工学専攻修了. (株) ドラゴンジェノミクス (現・タカラバイオ), 三共株式会社 (現・第一三共) 勤務を経て, 2008 年 奈良先端科学技術大学院大学情報科学研究科 (自然言語処理) 博士後期課程修了. 現在 奈良先端科学技術大学院大学情報科学研究科研究員. 博士 (工学).

新保 仁 (正会員)

1992 年 京都大学工学部電気工学第二学科卒. 1994 年 京都大学大学院工学研究科修士課程電気工学第二専攻修了. 1997 年 京都大学大学院工学研究科博士後期課程情報工学専攻指導認定退学. 博士 (工学). 現在 奈良先端科学技術大学院大学情報科学研究科助教. ACL, ACM, 日本データベース学会各会員.

松本 裕治 (正会員)

1977 年 京都大学工学部情報工学科卒. 1979 年 同大学院工学研究科修士課程情報工学専攻修了. 同年電子技術総合研究所入所. 1984–85 年 英国インペリアルカレッジ客員研究員. 1985–87 年 (財) 新世代コンピュータ技術開発機構に出向. 京都大学助教授を経て, 1993 年より奈良先端科学技術大学院大学教授, 現在に至る. 工学博士. 専門は自然言語処理. 情報処理学会, 日本ソフトウェア科学会, 言語処理学会, 認知科学会, AAAI, ACL, ACM 各会員.