

# 点予測による述語項構造解析

吉野 幸一郎<sup>1</sup> 森 信介<sup>1</sup> 河原 達也<sup>1</sup>

概要：本稿では点予測に基づく意味ラベル付与を用いた述語項構造解析について述べる。本手法では、意味ラベル付与器の学習に部分的アノテーションコーパスを用いることで、言語資源を最大限に活用することを目標とする。既存の述語項構造解析は係り受け構造を前提としているが、係り受け構造のフルアノテーションは高いアノテーションコストを要求する。そこでこうした情報を参照することなく、部分的アノテーションを許容する述語項構造解析の枠組みを設計し、係り受け木のアノテーションがないコーパスから直接意味ラベルの推定を行う。これにより、解析器のドメイン適応が容易になり、ドメインの移植性が向上する。部分的アノテーションによる学習を、意味ラベル付与器の学習において可能にするには、係り受け構造の情報を参照しないで学習を行う必要がある。そこで、係り受け構造以外の情報から点予測により意味ラベル付与器を学習し、評価を行った。その結果、効率的な学習と、従来の述語項構造解析器と同等の性能が達成されることが確認された。

## 1. はじめに

述語項構造とは、自然言語における最も基本的な意味表現の単位の1つである。質問応答システム [1]、テキストマイニング [2]、音声対話システム [3] など多くのアプリケーションシステムにおいて述語項構造が利用されており、述語項構造解析は自然言語処理における重要なタスクの1つであると言える。述語項構造は述語に対する格要素とその意味表現（意味役割ラベル）を表す構造である。例えば Propbank [4] においては、述語に係る枝に対して意味役割ラベルを付与することで、述語項構造を表現している。既存の意味役割解析 [5] においては、意味表現だけでなく、単語の係り受け構造が付与された学習データを利用して解析器の学習を行うことが一般的であった。

自然言語処理の上位のタスクになるに従って、一定量の文章に含まれる言語現象の数量は減少する。例えば1文書における述語項構造の数量は、係り受けの数量よりも小さい。既存の述語項構造解析手法は、いずれもこうした下位の現象（単語境界、品詞タグ、係り受け）のアノテーションが完全に行われていることを想定している。こうしたアノテーションが完全に行われているコーパスをフルアノテーションコーパスと呼ぶ。特に、学習データ中の全ての単語に対し、その係り先がどの単語であるかを、述語項構造解析の付与以前に指定する必要がある。言い換えると、述語

項構造解析のための学習データを1つ作るために、文書中の全ての単語の係り受け関係を付与しなければならないという問題がある。既存のフルアノテーションコーパスを利用する手法では低次の言語現象まで含めてアノテーションを必要とする制約があり、学習データを効率的に増やすことが難しい。加えて、係り受け関係のアノテーションには言語学上の専門知識を必要とする。このため、自然言語処理における上位の構造解析では、コーパスを容易に増やすことができず、ドメイン移植性を向上させることが難しいという問題がある。4.2節でも示すように、学習データの増加に従って述語項構造解析の精度は線形に向上し続けており、未だ上界に達していない。これは、学習データを増やすことでさらに精度向上が見込めるということである。重要な情報構造はドメインごとに異なることが指摘されており [6]、述語項構造解析の精度向上の観点からもドメイン移植は容易である必要がある。しかし、現実的な言語資源の状況を鑑みると、一般分野のフルアノテーションコーパスと対象分野の生コーパスが利用可能であることがほとんどであり、フルアノテーションコーパスのみを利用する既存手法ではドメイン移植性は著しく低くなってしまふ。このため、述語項構造解析においてより効率的な言語資源の活用が可能な枠組みが必要とされている。

そこで本論文では、点予測 [7] による述語項構造解析を提案する。形態素解析 [8] や係り受け解析 [9] においては、この枠組みによって効率的な言語資源の利用が可能となっている。点予測では、タスクに関連した部分だけのアノテーションを参照する。この枠組みではドメイン適応のた

<sup>1</sup> 京都大学 情報学研究所  
606-8501, 京都市左京区吉田本町  
Kyoto University, School of Informatics  
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

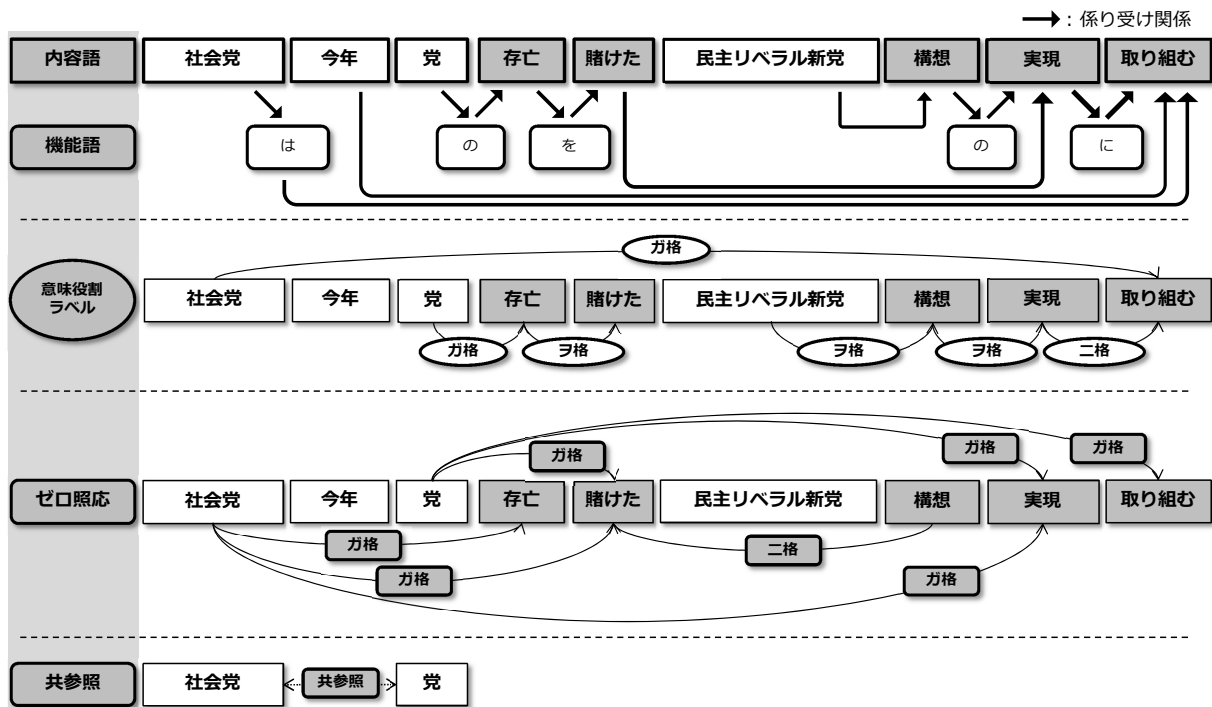


図 1 述語項構造解析の例 (灰色部分は述語)

めに、ドメイン固有の現象が生じている部分に対してのみ部分的アノテーションを行い、それを利用することが可能となる。本論文では、この点予測の枠組みを述語項構造解析のタスクに適用し、部分的アノテーションコーパスを利用することができる、新しい述語項構造解析の手法を提案する。特に係り受け構造は、全体を考慮したアノテーションを行う必要があるため、形態素解析結果から直接述語項構造を推定する枠組みを設計する。

点予測の枠組みでは、述語項構造解析器で用いる素性選択が重要となる。部分的アノテーションを許すためには、分類器で利用される素性がそれぞれ独立であるように素性選択を行う必要がある。そこで、述語項構造解析の枠組みを、曖昧性のない観測値による素性によって実現し、上位の曖昧性があるような情報は、それを含む観測値によって間接的に利用する。

こうした点を踏まえつつ、点予測による述語項構造解析の枠組みを提案し、既存手法と同等かそれを上回る解析精度を実現した。また、容易な学習データの追加やドメイン適応が効果的であることを示すため、実験によって学習データ量と解析精度の関係やドメイン適応の効果について実験を行い、部分的アノテーションコーパスの利用によって効果的な解析精度向上が実現できることを確認した。

## 2. 述語項構造解析

この節では、まず述語項構造の概念について述べる。また、既存の機械学習による述語項構造解析手法と、その問題点について述べる。

### 2.1 述語項構造

述語項構造解析とは、文書  $D$  中の述語  $P$  が入力として与えられた際に、その述語  $P$  に対応する格要素  $A_1, A_2, \dots, A_n$  とその意味役割  $S_1, S_2, \dots, S_n$  を付与するタスクである。この例を図 1 に示す。まず内容語と機能語の系列からなる文が入力として与えられ、それぞれの語には係り受け先が付与されている。次に、それぞれの係り受け関係に、意味役割ラベルを付与する。

続けて、意味役割ラベル付与後の 1 つ目の処理としてゼロ照応解析を行う。ゼロ照応解析とは、係り受け関係のない述語と格要素候補の間に述語項としてのラベルを付与する処理である。直接の係り受け関係にはないが述語項の関係を持つ場合をゼロ照応と呼ぶ。例では、「社会党」と「賭けた」は直接の係り受け関係にはないが、「賭けた」のゼロ照応におけるガ格は「社会党」であるということが言える。そこで、図 1 において「社会党」が述語「賭けた」のゼロ照応のガ格であるというラベルを付与する。この推定は述語項構造解析における重要な処理となっており、特に述語項構造を利用するアプリケーションはこの構造を必要としている。最後の処理は、意味ラベル付与後のもう 1 つの処理である共参照関係の付与である。共参照関係とは、文中の語同士、または文中の語と文外の語が実世界において同一の実体を指している場合、その関係のことを言う。例中の「社会党」と「党」は実世界において同一の実体を指しており、いずれも同じ述語の同じ格として扱うことができる。

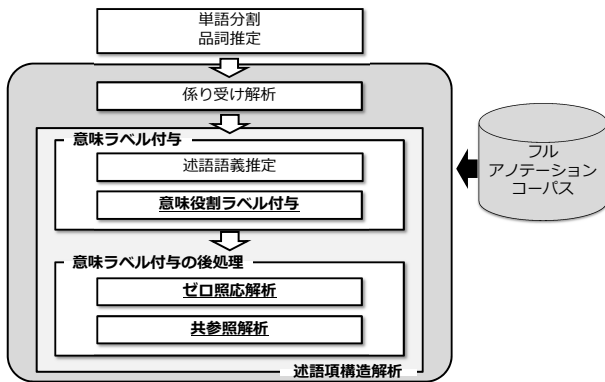


図 2 既存手法による述語項構造解析の概要

## 2.2 既存の述語項構造解析

既存の述語項構造解析の、最も一般的な解析手順を図 2 に示す。こうした解析器は、他の自然言語処理タスクと同様に機械学習に基づいて実現することができる。一般的な既存手法による述語項構造解析は、コーパスに対して必要な情報のアノテーションが行われていることを前提としている。述語項構造解析以前のアノテーション作業として、単語分割、品詞タグ付与、係り受け解析の 3 段階が必要である。その後意味役割ラベル付与を行うが、意味役割ラベル付与のタスクは 2 種類に分けることができる。1 つ目は述語語義の推定であり、2 つ目は意味役割ラベル付与である。これは、述語語義が意味役割ラベル付与に対して大きな影響を持つことに起因する。意味役割ラベル付与を行った後は、ゼロ照応解析と、共参照解析の 2 つのタスクを行う。

意味役割ラベルは係り受け関係と深い関係があることが知られており [10]、これらは係り受け関係の辺に付与されることが一般的であった。

## 2.3 意味役割ラベル付与後の処理

述語項構造解析は、ゼロ照応と共参照の問題を包含している。ゼロ照応の問題は、文書内で共有されている単語の省略によって起こり、前方照応によって解決できる [11]。この問題を解決するため、意味ラベル付与後の処理としていくつかの手法が検討されてきた [12], [13], [14]。これらの手法は *Salience Reference List* [11] と呼ばれる構造を利用して格要素候補の最尤のものを決定していたが、その結果、共参照という別の問題が引き起こされていた。

## 2.4 既存手法の問題点

前節で説明した通り、既存の機械学習を用いた述語項構造解析手法は単語境界、品詞および係り受けが付与されたフルアノテーションコーパスが必要となる。しかし、こうしたアノテーションを、訓練されていないアノテータが行うことは難しい。特に、係り受け構造のアノテーションには熟練したアノテータを必要とする。また、一部の係り受

けのアノテーションは全文体の構造を考慮しなければならないので、アノテーションを付与することが非常に難しい。こうしたアノテーションの困難さは、言語資源の有効活用やドメイン移植性の観点から問題である。

単語境界と品詞の推定に対しては、点予測によるアプローチ [8] によって効果的に言語資源を活用することで、高精度と高いドメイン移植性を両立できることが示されている。この手法を利用すれば、単語境界と品詞の推定においては、ドメイン移植を行う場合でも高い精度を実現できる。つまり、単語境界と品詞タグは、自動付与された結果をその後の処理に利用することが可能である。しかし、自動係り受け解析では、単語境界推定や品詞タグ付与のタスクと比較して高い精度を得ることが難しい。文献 [9] では点予測による効率的な言語資源の利用を係り受け解析において試行しているが、その解析精度の上界は決して高くはなく、自動の推定結果をそのままその後の処理に利用できるほどではない。このため、現実的な述語項構造解析において係り受けの素性を用いることは難しい。述語項構造のような意味情報を利用するアプリケーションシステムはドメイン適応が容易な枠組みを必要としている [3] が、係り受けの利用によってこれが困難になると言える。

しかし先述のように、単語境界と品詞タグは適応した自動推定結果を利用することが容易である。意味ラベル付与器をこうしたコーパスから学習することが可能になれば、ドメイン適応を容易にし、言語、人的双方の資源を有効に活用することができる。意味ラベル付与器を係り受けのアノテーションを用いずに学習する試みは既に存在しており [15]、この手法では推定された係り受け候補の森を利用することで、解析誤りの問題を軽減している。これは言語資源を有効活用するための方向性の 1 つである。しかし、この枠組みでは依然として係り受け解析器の適応が必要である。そこで、適応をより容易にしアノテーションコストを削減するため、係り受け解析器の適応を行わずに意味ラベル付与を行う枠組みを検討する。

## 3. 点予測による述語項構造解析

本稿では、適応を容易にするために係り受け解析と述語語義の推定を行わずに、述語項構造解析を行う。提案手法の概観を図 3 に示す。述語項構造解析の手法を再構築し、2 つの処理によって点予測による述語項構造解析を可能にする。提案手法は係り受けを利用する既存手法と異なり、単語境界と品詞タグのみが付与されたコーパスから直接述語項構造の解析を行う。

### 3.1 点予測

点予測とは文献 [7] で提案されたコンセプトで、全ての分類問題をそれぞれの分類点において、他の分類結果とは独立に解くというものである。系列ラベリング問題におい

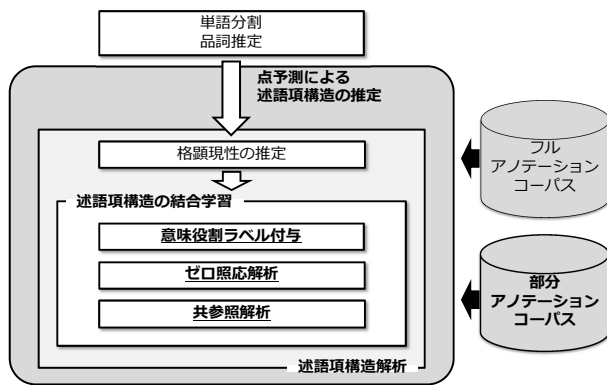


図 3 提案手法による述語項構造解析の概要

て、それぞれのラベルがお互いに影響を持っていることは事実であるが、その関係性を素性として用いる場合、推定結果を素性として用いることになり、その推定結果は真のラベルと同等に信頼性があるわけではない。そこで点予測は、推定値を素性として用いる代わりに、入力として与えられた真に信頼できる情報から、こうした推定値による特徴を間接的に参照する。これにより、部分的には整合性のない結果を出力するかもしれないが、全体的には系列ラベリング問題として解く場合と遜色ない精度を出すことが可能となる。もう一点、点予測を使うことで得られる利点として、部分的アノテーションコーパスからの学習が挙げられる。これは、点予測が他の分類問題の推定結果やラベルを参照しないことによる。

既存の述語項構造解析では、意味役割ラベルとゼロ照応のラベルはお互いのラベルに影響を与えるモデルを採用しており、さらにそれらは、係り受け構造と強い関係を持つという前提があった。しかしこうしたモデルは、全てのラベルが完全に付与されたフルアノテーションコーパスを用意しなければならず、言語資源の有効な活用が難しい。そこで点予測を用いるのだが、述語項構造解析のラベル付与問題を各ラベルごとに独立に解くためには、モデルの学習に周囲のラベル情報を参照せず、単独のラベル情報から学習できるようにする必要がある。さらに、部分的アノテーションコーパスを利用するためには、各ラベルの独立性を担保することも必要である。

### 3.2 格顕現性の推定

述語の種類によってどの格が付与されるか、また付与されやすいかといったことは異なる。例えばガ格はどの述語でもほぼ必須格となるが、ヲ格は述語の種類によって付与されない場合がある。これを格顕現性と呼び、まずこの推定を行う。これは格顕現性が、後段の意味ラベル付与に強く作用することが考えられるためである。

こうした問題を解決するために、どの述語にどのような格が出現しやすいかを捉える格フレームというものが提案されている [16]。格フレームによる格顕現性は既存の研究

でも利用されており、特に述語項構造解析の精度に大きく寄与するということが知られている [17]。そこで、本論文では格フレームによって格顕現性が与えられるものとして、この情報を意味ラベル付与の素性として用いる。例えば、図 1 における「賭ける」という述語は、ゼロ照応のガ格と係り受け関係にあるヲ格の 2 種類の格要素を持つことがわかる。既存研究では、これらは述語語義の推定によって解決されていた [18] が、これにはタスク固有のアノテーションが行われた言語資源が必要となる。こうした新たなアノテーションを必要とするような手法は点予測のコンセプトに反し、ドメイン移植性を低下させるため今回は用いない。

### 3.3 点予測による意味役割ラベル付与

次に、点予測による意味役割ラベル付与を行う。既存手法では、単語同士の係り受け関係が与えられた上で、係り受けの辺に対して意味役割ラベルの付与を行っていた [18]。しかし、こうした手法は係り受け関係のフルアノテーションコーパスを前提とし、多大なアノテーションコストが必要となる。単にアノテーションにかかるコストの他に、アノテータの訓練に必要となるコストも無視できない。

そこでこの問題を、より直接的な問題設計によって解決する。提案手法では意味役割ラベルの付与を、述語と格要素候補のペアに対する二値分類問題として行う。この問題の解法の例を図 4 に示す。フルアノテーションコーパスに対しては、意味ラベルが付与された述語と格要素のペアを正例とし、アノテーションが行われていないものを負例として学習を行う。この解析器の学習は 1 vs. rest によって行う。図 4 の例を取ってみると、述語「存亡」と格要素候補となりうる全ての単語の組み合わせから学習データを作り、ラベルが付与されているものは正例 (Y)、付与されていないものは負例 (N) とする。

部分的アノテーションコーパスの利用に際しては、述語と格要素候補のペアに対して状態を 3 つ定義する。1 つ目は正例、2 つ目は負例、3 つ目はラベルなしである。部分的アノテーションコーパスでは、ラベルが付与されていない事例は曖昧性があり、そのまま負例に利用することはできない。このため、アノテーションされていないドメインコーパスのうち、ドメイン固有で解析に影響があると考えられる点についてだけ正例か負例かのアノテーションを行う。図 4 の例では、「ヤンキース」と「勝った」のペアはゼロ照応のガ格の関係にあること (正例) と、「マリナーズ」と「浮上」がガ格の関係ではないこと (負例) という 2 つのドメイン固有の問題についてアノテーションが行われている。負例はフルアノテーションコーパス中で非常に多く出現するため、ドメインコーパスにおいて負例をアノテーションしなくても致命的な問題とはならない。アノテータは、ドメイン固有で解析に影響を及ぼしそうな事例を発見

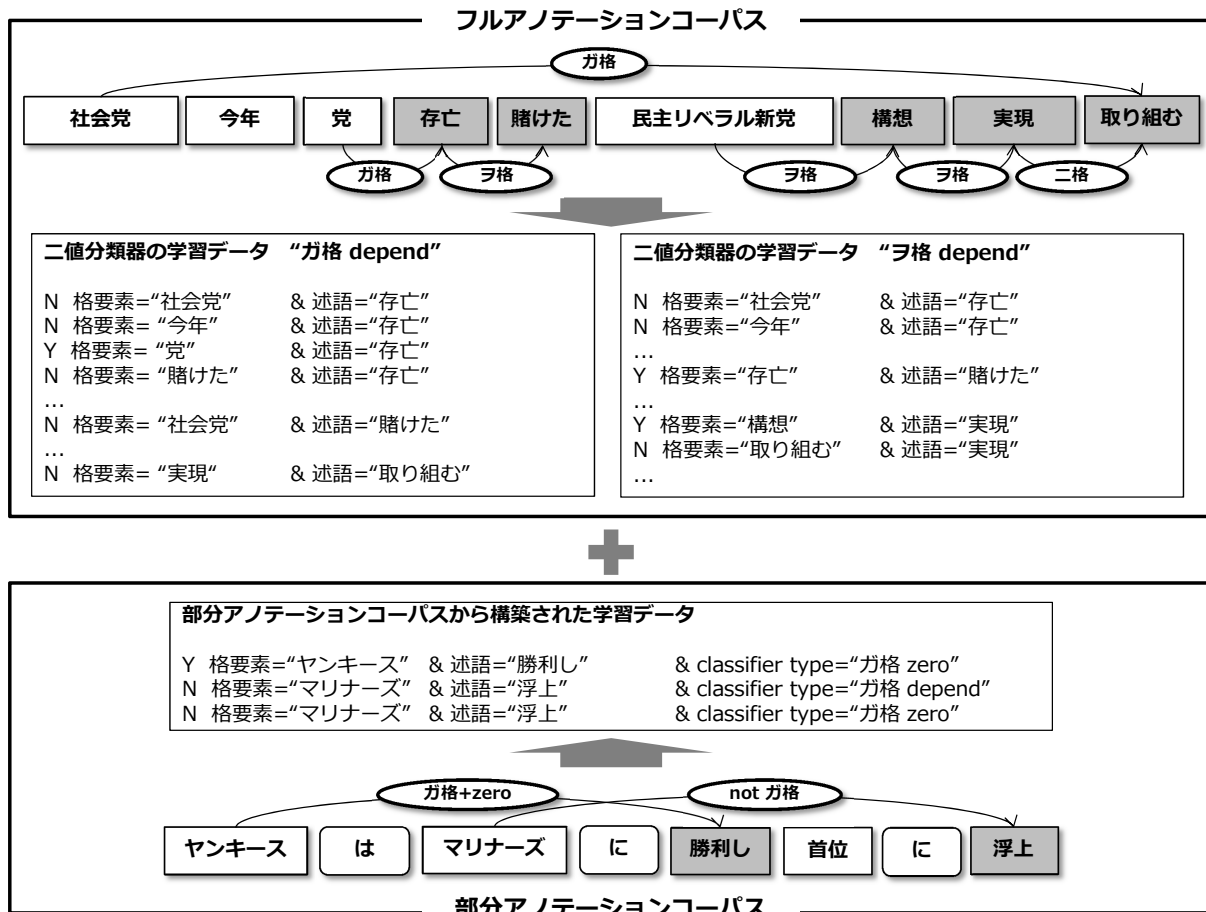


図 4 学習データの作成例

表 1 分類に利用する素性 ( $w_p$  は述語、 $w_a$  は格要素候補、 $w_{n-1}$  と  $w_{n-2}$  は各語の左側の語、 $w_{n+1}$  と  $w_{n+2}$  は各語の右側の語、 $t(w)$  は単語  $w$  の品詞)

種類	素性
単語 <i>uni-gram</i>	$w_{p-3}, w_{p-2}, w_{p-1}, w_p, w_{p+1}, w_{p+2}, w_{p+3}$ $w_{a-3}, w_{a-2}, w_{a-1}, w_a, w_{a+1}, w_{a+2}, w_{a+3}$
品詞 <i>uni-gram</i>	$t(w_{p-3}), t(w_{p-2}), t(w_{p-1}), t(w_p), t(w_{p+1}), t(w_{p+2}), t(w_{p+3})$ $t(w_{a-3}), t(w_{a-2}), t(w_{a-1}), t(w_a), t(w_{a+1}), t(w_{a+2}), t(w_{a+3})$
単語 <i>bi-gram</i>	$w_{p-1} + w_p, w_p + w_{p+1}, w_{a-1} + w_a, w_a + w_{a+1}$
品詞 <i>bi-gram</i>	$t(w_{p-1}) + t(w_p), t(w_p) + t(w_{p+1}), t(w_{a-1}) + t(w_a), t(w_a) + t(w_{a+1})$
単語 <i>tri-gram</i>	$w_{p-1} + w_p + w_{p+1}, w_{a-1} + w_a + w_{a+1}$
品詞 <i>tri-gram</i>	$t(w_{p-1}) + t(w_p) + t(w_{p+1}), t(w_{a-1}) + t(w_a) + t(w_{a+1})$
組み合わせ	位置 -2 - +2 の品詞の組み合わせ 述語と格要素候補の単語の組み合わせ
距離	述語と格要素候補の間にある述語の数
バイナリ	(1) 格助詞「は」「が」「を」「に」を左に持つ最も述語に近い格要素候補かどうか (2) 格助詞「は」「が」「を」「に」を左に持つ最も述語から遠い格要素候補かどうか (3) 格顕現性素性

した場合のみ、負例としてのアノテーションを行えば良い。ここからの学習データの生成についても、図 4 にその例が示されている。

分類問題にどのような素性を用いるかを表 1 に示す。まず、述語および格要素候補の単語と品詞、及び前後の単語と品詞からなる単純な  $n$ -gram の素性を用いる。また、前

後 2 個 (窓幅 5) の品詞については、組み合わせ素性を用いる。単語の組み合わせについては、スパースネスの問題から当該単語同士の組み合わせのみを用いる。格位置の分布は格の種類により異なるため、述語と格要素候補の距離も素性として用いる。距離の定義は、述語と格要素候補の間にある述語の個数とする。最後に、以下のバイナリ素

性を用いる。バイナリ素性 (1) と (2) は、センタリング理論 [19] に基づく素性である。Salience Reference List モデル [20] を述語項構造解析に適用した例 [21] でも、センタリング理論は格のスロットとして用いられている。Salience Reference List モデルでは、ガ格はしばしば省略されるという傾向がモデル化されている。これは、文書の最初の方にある格要素候補が、ガ格となりやすいことを示している。一方で、ヲ格や二格は省略されることが少なく、述語に近い格要素候補がヲ格や二格になりやすいということが言える。センタリング理論ではこうした文書全体の傾向を、係り受け構造上でモデル化している。点予測によるアプローチでこれと同等の素性を利用するため、独立な素性として定義できる上記のバイナリ素性を利用する。最後に、3.2 節で示される格顕現性を、バイナリ素性 (3) として用いる。

### 3.4 ゼロ照応解析と共参照解析

3.3 節で提案した二値分類によるタグ付与は、全ての格要素候補と述語の組に対して行われる。この手法をゼロ照応解析にも適用することができる。まず、格フレームから、「係り受け関係のガ格」、「ゼロ照応のガ格」、「係り受け関係のヲ格」、「ゼロ照応のヲ格」、「係り受け関係の二格」、「ゼロ照応の二格」が述語に対して存在するかのスロットを作る。それらのスロットそれぞれに対して、提案した二値分類問題の分類器を作成し、その分類器によって全ての述語と格要素候補の組み合わせを分類する。

この提案モデルでは全てのタグ付与問題を独立に解くことができるので、共参照もこの性質を利用して解くことができる。意味表現タグは独立に付与されるので、同一の述語に対して 2 つ以上の同じ意味表現タグを持つ格要素が存在する可能性があるが、それらを共参照関係がある単語として扱うことができる。つまり同じ分類器で正ラベルに分類された、同一の述語を持つ格要素候補が共参照となる。

まとめると、提案手法では点予測の性質を利用して、ゼロ照応や共参照の問題を同時に解くことができる。

### 3.5 述語項構造解析のドメイン適応

述語項構造を対話システムや情報抽出などの実アプリで利用する場合、対象とするドメインへの適応が重要な課題となる。ドメイン適応を行う場合、一般ドメインのフルアノテーションコーパスと適応先ドメインの生コーパスが利用可能である場合が多い。この中で、アノテーションされていないコーパスを最初からアノテーションするのは、現実的ではない。既存の係り受けの上に述語項構造を付与する手法に沿うと、係り受けと意味役割ラベルの両方を付与したフルアノテーションコーパスを用意する必要がある。述語項構造解析のドメイン適応を行う際に重要なのは、ドメイン依存の重要単語とその周辺のアノテーションであるが、フルアノテーションコーパスのみしか利用できない枠

表 2 コーパスの諸元

コーパス	NTC_train	NTC_adapt
文書数	2,318	609
文数	19,541	18,711
述語数	90,302	66,934
ガ格 depend	27,179	23,675
ヲ格 depend	20,763	19,825
二格 depend	9,474	547
ガ格 zero	130,852	103,481
ヲ格 zero	15,177	9,660
二格 zero	12,247	450

組みにおいては、それ以外の一般ドメインコーパスによって被覆することができる部分についてもアノテーションを行わなければならない。部分的アノテーションに関する既存研究でも、こうした問題は提起されている [22], [23]。部分的アノテーションを用いる場合、これまでの枠組みでは扱うことができなかった部分的にタグが付与されていないコーパスを許容する必要がある。

点予測では分類器は系列的な素性や推定値による素性を学習に用いず、確定的な情報のみを用いる。この特長により、点予測によって構築された分類器は部分的アノテーションコーパスの確定的な情報のみを利用することができる。既存手法で用いられる系列的な素性や構造的な素性によって部分的アノテーションコーパスから学習を行おうとすると、アノテーションの不足などから、必要な条件付き確率を全て計算することが難しい。しかし点予測では各分類問題を独立に扱うので、こうした部分アノテーションコーパスの利用が可能となる。こうした特長は、分類器のドメイン適応・再学習を容易にする。

## 4. 実験

提案手法を評価するため、述語項構造解析の精度、コーパスサイズによる解析精度の変化、ドメイン適応の効果という 3 種類の実験を行った。実験には京都テキストコーパス version 4.0 [24] をベースとして、述語項構造と共参照関係をアノテーションすることによって構築された NAIST テキストコーパス (NTC) [25] を用いた。NTC は日本語の新聞記事 (毎日新聞) に対して付与されており、新聞記事と社説の 2 種類のドメインを含む。述語項構造解析を新聞記事 (NTC\_train) の中で評価し、社説記事 (NTC\_adapt) をドメイン適応先として用いた上でドメイン適応の評価を行う。それぞれの評価は 5 分割交差検定により行う。実験に利用されたコーパスの詳しい諸元を表 2 に示す。

NTC ではガ格 (主格) ヲ格 (直接目的格) 二格 (間接目的格) の 3 種類のタグが、述語に対してペアとなる格要素の単語と併記して付与されている。全てのタグは depend (係り受け関係) か zero (ゼロ照応) の属性を持っており、述語に対して格要素が直接係っているか、ゼロ照応として

表 3 ラベルの付与精度 (格顕現性素性なし)

属性	格	適合率	再現率	F 値
depend	ガ格	0.6097	0.5422	0.5740
	ヲ格	0.8682	0.8152	0.8409
	二格	0.8316	0.7989	0.8149
	平均	0.7435	0.6833	0.7121
zero	ガ格	0.2300	0.1109	0.1496
	ヲ格	0.4396	0.1698	0.2450
	二格	0.3094	0.0842	0.1324
	平均	0.2506	0.1145	0.1572

表 4 意味役割ラベル付与の精度 (格顕現性素性あり)

手法	格	適合率	再現率	F 値
提案手法 (depend 属性)	ガ格	0.7468	0.7536	0.7502
	ヲ格	0.9082	0.9300	0.9190
	二格	0.9530	0.9474	0.9502
	平均	0.8393	0.8494	0.8443
比較手法 [5]		-	-	0.7869

表 5 ゼロ照応解析の精度 (格顕現性素性あり)

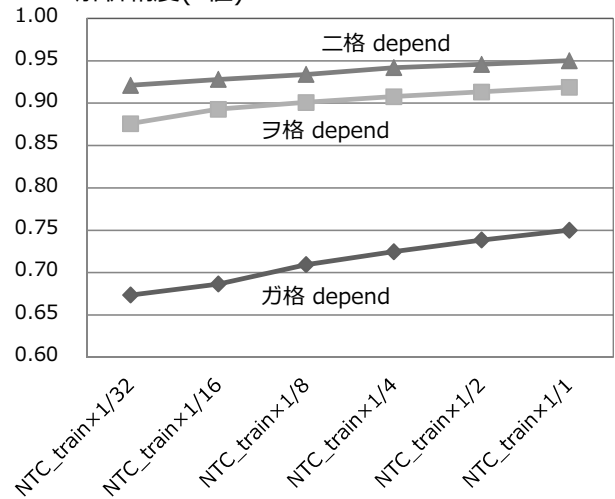
手法	格	適合率	再現率	F 値
提案手法 (zero 属性)	ガ格	0.3046	0.1201	0.1723
	ヲ格	0.5598	0.2116	<b>0.3070</b>
	二格	0.4015	0.1265	<b>0.1924</b>
	平均	0.4015	0.1265	0.1924
比較手法 [26]	ガ格	0.2646	0.3015	<b>0.2819</b>
	ヲ格	0.0918	0.1291	0.1073
	二格	0.0475	0.0405	0.0437
	平均	0.2308	0.2658	0.2469

係っているかがわかるようになっている。述語項構造より下位のタスクにおいては、単語境界、品詞タグ、文節単位の係り受けが付与されている。これらの中から、単語境界と品詞タグはそのまま利用してガ格-depend、ガ格-zero、ヲ格-depend、ヲ格-zero、二格-depend、二格-zero の 6 種類の分類器を構築した。

#### 4.1 述語項構造解析の精度

提案した点予測による述語項構造解析器を、表 1 で説明した素性を用いて LIBLINEAR (線形 SVM) [27] によって学習した。評価は付与された意味役割ラベルの適合率、再現率とそれらの調和平均 (F 値) によって行う。5 分割交差検定による評価結果を表 3、表 4、表 5 に示す。表 3 に示すのは、格顕現性の素性 (表 1 でのバイナリ素性 (3)) を用いない場合の解析精度である。表 4、表 5 で示した格顕現性の素性を用いた場合の結果と比較すると精度の大きな低下が見られ、格顕現性の推定が重要なタスクであることがわかる。表 4 では、depend 属性のタグ付与精度を先行研究 [5] と比較する。先行研究は主格、直接目的格、間接目的格以外にもラベルを含むので正確な比較ではないが、提案手法はおおよそ先行研究と同等の解析精度を実現でき

解析精度 (F 値)



解析精度 (F 値)

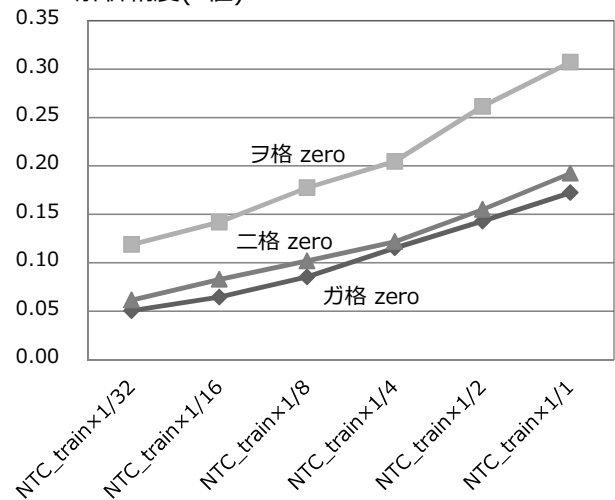


図 5 コーパスサイズと解析精度

ていることがわかる。表 5 ではゼロ照応解析の精度を、先行研究 [26] と比較する。先行研究ではゼロ照応解析を文内と文間で分けて評価しているの、比較のために加重平均を取った値を表に示す。ゼロ照応解析の精度でも、先行研究と同等の解析精度を実現できており、特にヲ格と二格では先行研究の精度を上回っていることがわかる。この結果により、提案した点予測を用いた述語項構造解析が、既存の述語項構造解析と同等に動作することが確認された。点予測では既存手法よりも単純な素性が用いられているが、素性選択の工夫により複雑な素性を用いた既存手法と同等の精度を実現可能であることがわかる。

#### 4.2 コーパスサイズと解析精度

コーパスサイズと解析精度の関係を図 5 に示す。横軸は対数スケールのコーパスサイズ (学習に利用したコーパスサイズ) であり、縦軸は解析精度の F 値である。評価は 5 交差検定により行い、4 つある学習データの使用量を変化させた。図から解析精度はコーパスサイズに従い線形に向

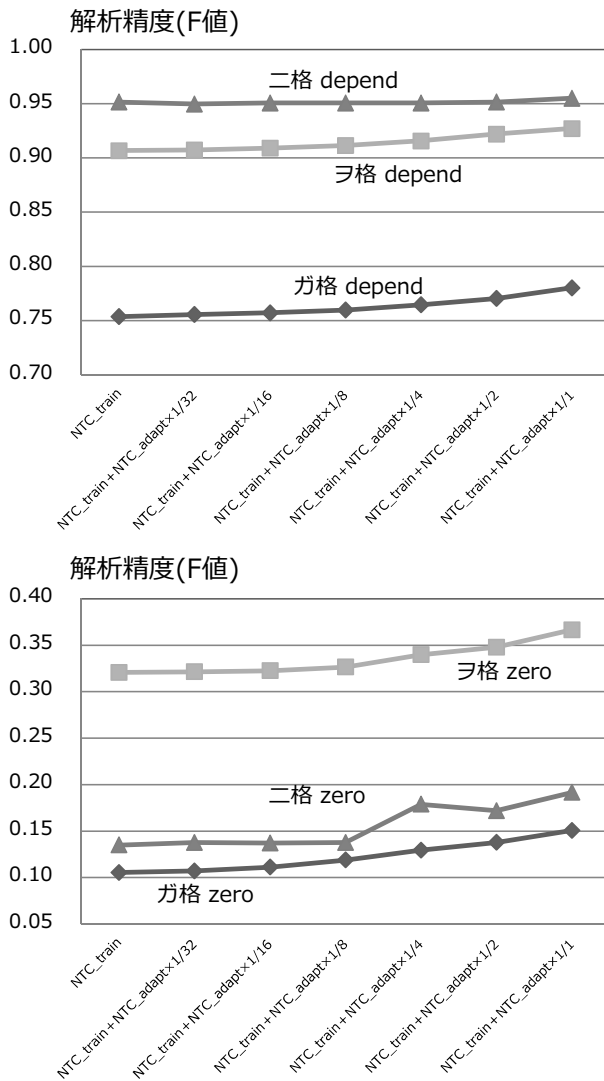


図 6 ドメイン適応の効果

上しており、利用可能な学習データ全てを用いてもまだ収束していないことがわかる。これはコーパスの効果的なアノテーションと利用が精度に大きく寄与することを示している。

#### 4.3 ドメイン適応の評価

ドメイン適応の効果に関する評価実験の結果を図 6 に示す。一般ドメインの全てを学習データに加えた上で、適応先ドメインのコーパスの 5 交差検定を行った。つまり、学習には 5 等分された適応先ドメインコーパスの 4 個と一般ドメインコーパスの全てを用い、残った 1 個をテストデータとすることを 5 回繰り返した。適応先ドメインコーパスの量による効果を示すため、学習に用いたドメインコーパスの 4 個をさらに細分し、学習に用いた。横軸はこれによって得られる、適応先ドメインコーパスの量を対数スケールで表したものである。このグラフにより、提案した枠組みによりドメイン適応が上手く行えることが確認され、点予測による枠組みが解析精度を向上させることが確

認できた。zero の二格においてのみデータサイズに従っていないように見られるが、これは今回の適応先ドメインにおいて二格が極端に少ないことによる。今回用いた新聞の社説記事においては、609 文書の中で直接係り受け関係にある二格が 547 個、ゼロ照応としての二格が 450 個しか見られなかった。

#### 5. まとめ

点予測を用いた述語項構造解析の枠組みを提案し、この学習に部分的アノテーションコーパスが利用可能であることを示した。この結果、提案手法は言語資源を構築する手間を削減し、述語項構造解析を様々なドメインに対して適応することを可能とした。述語項構造解析を各個に独立した単純な素性を用いて行うことにより、部分的アノテーションコーパスを利用することが可能になる。実験によって、この枠組みで既存の述語項構造解析と同等の解析精度を実現でき、精度向上のための言語資源の追加を容易に行うことができることを示した。

我々の実験によれば、述語項構造解析の精度はコーパスサイズの増加に従って向上している。この傾向は一般ドメイン (表 5) でも適応先ドメイン (表 6) でも見られ、効果的な学習データの追加によって解析精度が向上し続けることを示している。これはつまり、必要な部分だけをアノテーションした部分的アノテーションコーパスを利用することによって、述語項構造解析精度をより向上させることができることを示している。

この枠組みを活用するため、効果的なアノテーションを行う仕組みの開発が必要となる。また、他の有効な素性はこういったものか、推定値を素性として追加する場合はどのような枠組みを加えればよいかといった検討も必要である。

#### 参考文献

- [1] Shen, D. and Lapata, M.: Using Semantic Roles to Improve Question Answering, *Proc. of EMNLP-CoNLL*, pp. 12–21 (2007).
- [2] Wang, R. and Zhang, Y.: Recognizing Textual Relatedness with Predicate-Argument Structure, *Proc. of EMNLP*, pp. 784–792 (2009).
- [3] Yoshino, K., Mori, S. and Kawahara, T.: Spoken Dialogue System based on Information Extraction using Similarity of Predicate Argument Structures, *Proc. of SIGDIAL*, pp. 59–66 (2011).
- [4] Palmer, M., Gildea, D. and Kingsbury, P.: The proposition bank: An annotated corpus of semantic roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106 (2005).
- [5] Watanabe, Y., Asahara, M. and Matsumoto, Y.: A Structured Model for Joint Learning of Argument Roles and Predicate Senses, *Proc. of ACL*, pp. 98–102 (2010).
- [6] R.Grishman: Discovery Methods for Information Extraction, *Proc. of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 243–247 (2003).



- [7] Neubig, G. and Mori, S.: Word-based partial annotation for efficient corpus construction, *Proc. of LREC* (2010).
- [8] Neubig, G., Nakata, Y. and Mori, S.: Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, *Proc. of ACL*, pp. 529–533 (2011).
- [9] Flannery, D., Miyao, Y., Neubig, G. and Mori, S.: Training Dependency Parsers from Partially Annotated Corpora, *Proc. of IJCNLP*, pp. 776–784 (2011).
- [10] Johansson, R. and Nugues, P.: Dependency-based semantic role labeling of PropBank, *Proc. of EMNLP*, pp. 69–78 (2008).
- [11] Iida, R., Inui, K. and Matsumoto, Y.: Zero-Anaphora Resolution by Learning Rich Syntactic Pattern Features, *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 6, No. 4, pp. 12:1–12:22 (2007).
- [12] Sasano, R. and Kurohashi, S.: A Probabilistic Model for Associative Anaphora Resolution, *Proc. of EMNLP*, pp. 1455–1464 (2009).
- [13] Iida, R. and Poesio, M.: A cross-lingual ILP solution to zero anaphora resolution, *Proc. of ACL-HLT*, pp. 804–813 (2011).
- [14] Hayashibe, Y., Komachi, M. and Matsumoto, Y.: Japanese Predicate Argument Structure Analysis Exploiting Argument Position and Type, *Proc. of IJCNLP*, pp. 201–209 (2011).
- [15] Boxwell, S. A., Brew, C., Baldrige, J., Mehay, D. and Ravi, S.: Semantic Role Labeling Without Treebanks?, *Proc. of IJCNLP*, pp. 192–200 (2011).
- [16] Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, *Proc. of HLT-NACCL*, pp. 176–183 (2006).
- [17] Sasano, R., Kawahara, D. and Kurohashi, S.: A fully-lexicalized probabilistic model for Japanese zero anaphora resolution, *Proc. of COLING*, pp. 769–776 (2008).
- [18] Björkelund, A., Hafdell, L. and Nugues, P.: Multilingual semantic role labeling, *Proc. of CoNLL: Shared Task*, pp. 43–48 (2009).
- [19] Grosz, B. J., Weinstein, S. and Joshi, A. K.: Centering: a framework for modeling the local coherence of discourse, *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225 (1995).
- [20] Nariyama, S.: Grammar for ellipsis resolution in Japanese, *Proc. of TMI*, pp. 135–145 (2002).
- [21] Iida, R., Inui, K., Takamura, H. and Matsumoto, Y.: Incorporating contextual cues in trainable models for coreference resolution, *Proc. of EACL Workshop on the Computational Treatment of Anaphora*, pp. 23–30 (2003).
- [22] Tsuboi, Y., Kashima, H., Oda, H., Mori, S. and Matsumoto, Y.: Training conditional random fields using incomplete annotations, *Proc. of COLING*, pp. 897–904 (2008).
- [23] Sassano, M. and Kurohashi, S.: Using smaller constituents rather than sentences in active learning for Japanese dependency parsing, *Proc. of ACL*, pp. 356–365 (2010).
- [24] Kawahara, D., Kurohashi, S. and Hasida, K.: Construction of a Japanese relevance-tagged corpus, *Proc. of LREC*, pp. 2008–2013 (2002).
- [25] Iida, R., Komachi, M., Inui, K. and Matsumoto, Y.: Annotating a Japanese text corpus with predicate-argument and coreference relations, *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139 (2007).
- [26] 笹野遼平, 黒橋禎夫: 大規模格フレームを用いた識別モデルに基づく日本語ゼロ照応解析, *情報処理学会論文誌*, Vol. 52, No. 12, pp. 3328–3337 (2011).
- [27] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification, *Journal of Machine Learning Research*, Vol. 9, No. 4, pp. 1871–1874 (2008).