Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

文脈情報と格構造の類似度を用いた日本語文間述語項構造解析

林部祐太[†1]　小町守[†1]　松本裕治[†1]

文脈情報と格構造の類似度を用いた日本語文間述語項構造解析手法を提案する．
センタリング理論に基づく局所文脈情報と述語と項候補の共起頻度といった意味的情報という大まかには 2 つの情報を用いて従来の文間述語項構造解析は行われてきた．ところが，いずれの手法を用いても，「X を逮捕した」という文をもとに「自首した」のガ格項が X であると判定することはできなかった．そこで本論文では，格構造の類似度と述語項構造解析の履歴を用いることで，文章全体の文脈情報（大域文脈情報）から文間述語項構造解析を行うことを提案する．

Improving Japanese Inter-sentential Predicate Argument Structure Analysis with Contextual Information and Similarity between Case Structures

Yuta Hayashibe ,[†1] Mamoru Komachi [†1]
and Yuji Matsumoto [†1]

We improve Japanese inter-sentential predicate argument structure analysis with contextual information and similarity between case structures.
Two types of clues have been often used in previous work. One is local contextual information based on centering theory, and the other is semantic information such as co-occurrences between a predicate and an argument candidate. However, those approaches fail to identify the nominative argument in the sentence “He turned himself in to police”, even if the document has a sentence like “The police arrested him.” Thus, we propose a new method using global contextual information and similarity between case structures in order to exploit global contextual information over a document.

1. はじめに

> 太郎はカレーを持ってきた．
> すぐに次郎は食べた．

といった文章において,「食べた」のは「太郎」ではなく「次郎」で,「次郎」が食べたのは「カレー」である．このように，文章から「誰が何をどうした」という意味的な関係を抽出することを**述語項構造解析**といい，機械翻訳や自動要約などの自然言語処理の応用において重要である．

「次郎」や「カレー」のように述語との関係をもつ名詞句を**項**とよぶ．述語と同一文内にある項を文内項，述語と異なる文にある項を文間項という．解析に用いられる情報が少ないことや項候補数が多いことから，文内項より文間項の方が，解析は一般に難しい．本稿では，文間項の述語項構造解析を対象とする．

文間項構造解析では，共起情報やセンタリング理論をもとにした局所文脈情報，また項候補が以前の解析で他の動詞の項となった回数の情報といった文章全体の解析から得られる大域文脈情報の利用の有効性が報告されている．(詳細は 2 節で述べる)

ところが，このいずれを用いても単独では

> 府警は花子を逮捕した．
> (ϕ が) 昨日自首したようだ．

という文章の ϕ が何であるかを同定することはできない．

そこで本研究では，各項候補が以前の文で「どのような述語の項となったか」に着目して解析することを提案する．例えば，先の例文では,「府警」と「花子」は，それぞれ「逮捕した」のガ格・ヲ格であるが,「自首した」のガ格が「逮捕した」のヲ格と同じような項を持ちやすいという情報を用いられれば，正しく ϕ が「花子」であると解析することができる．

本研究では NAIST テキストコーパスを対象に文間述語項構造解析を行い，提案手法の効果を検証し，大域文脈情報を利用した他の手法との比較も行う．

†1 奈良先端科学技術大学院大学
Nara Institute of Science and Technology

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

2. 関 連 研 究

文間述語項構造解析では項候補を述語とは別の文から探すことから，文章全体の談話構造（文脈）を捉えることが重要である．

本節では，文脈の情報を方法として，着目する述語の直前の高々数文の情報を用いる局所文脈情報と，すべての文の情報を用いる大域文脈情報の関連研究について述べる．

2.1 局所文脈情報の利用

談話構造と話題の移り変わりを説明するセンタリング理論[1] をもとに，飯田らは Salience Reference List (SRL)[2] を用いた手法を提案している[3]．SRL はハ格・ガ格・ヲ格・ニ格の項候補を 1 つずつ保持する 4 つのスロットからなるリストで，次の手順で作成する．

(1) 文章の最初から順に各先行詞候補についてスロットに該当するかを調べる
(2) 該当する場合スロットに格納する
(3) すでにスロットに格納されている場合，上書きして格納する
(4) 調査する述語の直前まで繰り返す

飯田らは SRL をセンタリング理論に基づく日本語照応解析処理のモデルと同様に，

主題（ハ格）＞主語（ガ格）＞間接目的（ニ格）＞直接目的（ヲ格）＞その他

と順序付けて利用した．この選好は日本語ではどのようなものが省略されやすいかを示している．

次の例文と**表 1** を用いて,「開始した」のガ格の SRL による解析を示す．

ドゥダエフ大統領$_{(1)}$ は，正月休戦$_{(2)}$ を提案したが，エリツィン・ロシア大統領$_{(3)}$ はこれを黙殺し，(φが) 行動を開始した．

SRL は最初すべてのスロットが空である．まず (1) について調べ，(1) はハ格なのでハ格のスロットを埋める．次に (2) について調べ，(2) はヲ格なのでヲ格のスロットを埋める．最後に (3) について調べ，(3) はハ格なのでハ格のスロットを書き換える．

最終的にできた SRL によると最尤先行詞は「エリツィン・ロシア大統領」である．

2.2 大域文脈情報の利用

文章全体の文脈を捉える 1 つの方法として，各項候補が，どのくらい他の述語の項として使われたのかという情報の利用が考えられる．これは，センタリング理論の立場からも，統計的な観点から[4] も一度項になった名詞句は再び項になりやすいという知見に基づいている．

表 1 SRL による解析例
Table 1 A sample analysis with SRL

	最初	(1)	(2)	(3)
ハ	-	ドゥダエフ大統領	ドゥダエフ大統領	エリツィン・ロシア大統領
ガ	-	-	-	-
ヲ	-	-	正月休戦	正月休戦
ニ	-	-	-	-

表 2 各候補の CHAIN_LENGTH 素性の値
Table 2 The value of CHAIN_LENGTH feature of each candidate

	目指す$_{(a)}$	**進め**$_{(b)}$
A 社	2	3
記者会見	1	1
B 社	0	0
提携	1	1
共同開発	-	0
市場	-	0
開拓	-	1

※「-」は項候補にならないことを示す

2.2.1 CHAIN_LENGTH 素性

飯田らは，項候補が何回項となったかという情報を機械学習の素性 (CHAIN_LENGTH 素性) として用いた[5][3]．次の文章と**表 2** を用いて，ガ格の文間項同定の例を示す[⋆1]．なお，飯田らは評価実験時には，着目している述語の前方にある照応・省略表現の解析はすべて正しく行われていると仮定していた[6] ため，ここでもそれに倣う．

A 社 は 記者会見 を**開き**，B 社との 提携 を**発表した**．
共同開発により，新たな市場の 開拓$_{(a)}$．
昨秋から交渉を**進め**$_{(b)}$ ていたらしい。

まず述語**「目指す」**$_{(a)}$ について考える．項候補は「A 社」,「記者会見」,「B 社」,「提携」の 4 つである.「A 社」は**「開き」**と**「発表し」**の項になっているので値は 2,「記者会見」と「提携」はそれぞれ**「開き」**と**「発表し」**の項となっているので値は 1 となる．一方,「B 社」

⋆1 文章の先頭から動詞の項構造解析を行い，項探索範囲は着目する述語がある文の以前にある文章とする

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

表 3 本論文の実験設定における USED 素性の値
Table 3 The value of USED feature of each candidate under this paper's experimental setting

	目指す$_{(a)}$	進め$_{(b)}$
A 社	1	1
記者会見	1	1
B 社	0	0
提携	1	1
共同開発	-	0
市場	-	0
開拓	-	1
昨秋	-	-
交渉	-	-

表 4 Imamura らの実験設定における USED 素性の値
Table 4 The value of USED feature of each candidate under this paper's experimental setting

	目指す$_{(a)}$	進め$_{(b)}$
A 社	1	1
記者会見	1	1
B 社	-	-
提携	1	1
共同開発	0	-
市場	0	-
開拓	0	1
昨秋	-	0
交渉	-	0

※「-」は項候補にならないことを示す

はどの述語の項にもなっていないので，値は 0 となる．

次に，**「進め」**$_{(b)}$ について考える．項候補は「A 社」,「記者会見」,「B 社」,「提携」,「共同開発」,「市場」,「開拓」の 7 つである.「記者会見」と「共同開発」と「市場」は項となった回数が 0 回,「提携」と「提携」は 1 回,「A 社」は 3 回なので，CHAIN_LENGTH 素性の値はそれぞれ 0,1,3 となる．

2.2.2 USED 素性

Imamura らも飯田らと同様の観点から，項候補が以前に項として使われたかどうかという真偽値の情報を機械学習の素性（USED 素性）として用いた[7]．すなわち，USED 素性は，CHAIN_LENGTH 素性が 0 なら偽，0 でないならば真となる．例文より USED 素性を作成すると**表 3** のようになる．

なお，Imamura らの実験設定では文間項構造解析であっても述語の探索範囲は述語のある文も含めて解析を行っているが，平均項候補数は 102.2 と多いため項探索範囲を同一文内の項候補と以前の文で項になったものに限定している．すなわち，述語のある文より前の文に位置する名詞句のうち，USED 素性の値が 0 のものは項候補とはしていない．そのため，ほとんどの USED 素性は**表 4** のように 1 となる．

2.3 事態間の類似度

1 節で述べたような「〜が自首する」と「〜を逮捕した」という 2 つの事態について，それらが類似しているとどうかを捉える手法として，事態間の自己相互情報量 (PMI) を計算する手法が提案されている[8]．

表 5 SRL による解析例
Table 5 A sample analysis with SRL

ハ	府警
ガ	被害
ヲ	A
ニ	-

表 6 被使用情報
Table 6 "used" information of each candidate

	使われ方	CHAIN LENGTH 素性の値
府警	が逮捕する	1
A	を逮捕する	1
被害	が多発する	1
その他の候補	-	0

2 つの事態 e_1 と e_2 の類似度は次式で定義され，大量の文書を用いて計算できる．事態 e は述語と格から定まり，e_1,e_2 それぞれの述語を w,v，格を d,g とする．なお $C(e_1, e_2)$ は 2 つの事態 $e(x, d)$ $e(y, f)$ が d と f がコーパス中において照応関係になっている回数である．

$$pmi(e_1, e_2) = \log \frac{P(e_1, e_2)}{P(e_1)P(e_2)}$$

$$P(e_1, e_2) = \frac{C(e_1, e_2)}{\sum_{x,y} \sum_{d,f} C(e(x, d), e(y, f))}$$

ところがこの手法は，照応解析がある程度うまくいかなければ使えないことと，事態の共起を計算に用いているため相当大量の文書を収集しなければならないことが問題点である．

3. 大域文脈情報としての格構造の類似度の利用

次の文章の述語「**自首する**」のガ格の文間項同定を考える．(正解は「A」である)

> 府警は一連の窃盗事件の容疑で A を逮捕した．
> 最近 B 地区では被害が多発していた．
> 「逃げ切れないと思い**自首し**た」と供述している．

局所文脈情報として SRL を用いた解析では，**表 5** のようになり，この情報からは間違って「府警」が選択されてしまう．一方，大域文脈情報として他の述語の項として使われた回数を用いた解析では，**表 6** のようになり「府警」や「被害」と比べて「A」も同じく 1 回の使用なので，それらとは差がつかない．

ここで，**表 7** に示したそれぞれの項の分布を見ると,「が自首する」と「を逮捕する」は似たような項を伴っており，自首した人が逮捕する可能性よりは，自首した人が逮捕される可能性が高いことが予想できる．

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

表 7 各項構造の項分布（頻度順）
Table 7 Argument distributions of each case structure (Sorted by frequency)

格構造	1599	が自首する	5651	が逮捕する	82112	を逮捕する	69973	が多発する
	136	者	702	署員	15285	人	10449	事件
	117	犯人	698	警察	8484	者	6357	事故
	96	彼	376	警察官	5563	男	3377	犯罪
	68	人	368	署	2804	名	2289	トラブル
	63	男	230	県警	1188	犯人	2245	現象
	36	犯	177	員	1185	男性	2000	災害
項	30	少年	153	府警	763	ら	1861	こと
	26	高校生	137	当局	671	おまえ	1744	被害
	23	梶	132	警視庁	587	ところ	1579	ケース
	22	人間	127	人	562	氏	1543	問題
	21	女性	126	容疑	482	少年	1309	など
	20	息子	115	警官	449	人々	1111	地震

しかし，先に示したように 2 節で述べた文脈情報を文間述語項構造解析に用いる方法では，この傾向を捉えることはできない．そこで本研究ではこの傾向を統計的に扱うために，**格構造の類似度を用いた項構造解析履歴の利用**を提案する．

本節では，まず格構造の類似度を定義し，次に項構造解析への利用方法について述べる．

3.1 格構造の類似度の定義

本研究では「格構造」を助詞と述語の組と定義し，2 つの格構造が似たような項の分布を持つとき,「格構造が類似している」と定義する．

2 つの項分布 p, q の類似度 $Sim(p, q)$ は Jensen-Shannon divergence（以下 JS ダイバージェンスとよぶ）を用いて次のように定義する．なお類似度を計算する際は，それぞれの格構造において，項数の合計値で除算することで，項分布を正規化しておくとする．

$$
\begin{aligned}
Sim(p,q) &= 1 - JS(p,q) \\
KL(p,q) &= \sum p(x) \log \frac{p(x)}{q(x)} \\
&= -\sum p(x) \log q(x) + \sum p(x) \log p(x) \\
r(x) &= \frac{p(x)+q(x)}{2} \\
JS(p,q) &= \frac{1}{2}\left(KL(p,r) + KL(q,r)\right) \\
&= \frac{1}{2}\left(\sum p(x) \log \frac{p(x)}{\frac{p(x)+q(x)}{2}} + \sum q(x) \log \frac{q(x)}{\frac{p(x)+q(x)}{2}}\right)
\end{aligned}
$$

表 8 格構造類似度
Table 8 Similarities between case structures

	が自首する	が逮捕する	を逮捕する	が多発する
が自首する	1	0.4590	0.7568	0.3777
が逮捕する		1	0.4861	0.3654
を逮捕する			1	0.4044
が多発する				1

JS ダイバージェンスの性質より，左下の部分は対角線を介して右上の部分と線対称となるため省略した

JS ダイバージェンスは 2 つの分布の類似度を図る尺度で，この値が小さいほど分布が似ていることを示す．また次のような特徴をもつ．

- $0 \leq JS(p,q) \leq 1$
- $p = q$ のときかつそのときのみ $JS(p,q) = 0$
- $JS(p,q) = JS(q,p)$

このようにして求めた格構造類似度の例を**表 8** に示す.「が自首する」と「を逮捕する」の項分布の類似を捉えられていることが分かる．

なお本手法は 2.3 節で述べた手法と比べて，照応解析を行わなくても良い点と，1 つの文書から大量の共起を獲得できる点で優れている．

3.2 格構造と項構造解析履歴の類似度の定義

ある述語の格構造 p と，名詞句 n の項構造解析履歴 $H = \{h_1, h_2, \cdots, h_n\}$ の類似度 $Sim2(p, H)$ を次のとおり定義する．H は「〜を食べる」,「〜を買う」などの，名詞句 n が着目している述語以前に項となったものの集合である．なお，n と照応関係にある名詞句の項構造解析履歴も H に含むとする．

$$Sim2(p, H) = \max_i \left(Sim(p, h_i)\right)$$

例えば，名詞句「泥棒」が項構造解析履歴「を逮捕する，が自首する」を持つ場合，格構造「が多発する」と履歴の分布類似度は

$$
\begin{aligned}
&Sim2(\text{が多発する}, \{\text{ を逮捕する}, \text{が自首する }\}) \\
&= \max\left(\{Sim(\text{が多発する}, \text{を逮捕する}), Sim(\text{ が多発する}, \text{が自首する }\})\right) \\
&= \max\left(\{0.3654, 0.4044\}\right) \\
&= 0.4044
\end{aligned}
$$

となり，0.4044 と求まる．

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

3.3 文間項構造解析における利用

以上で定義した格構造と項構造解析履歴の類似度 $Sim2$ は文間項構造解析の手がかりとして用いることができる．

例えば，「自首した」のガ項の同定では，「府警」，「A」，「被害」のそれぞれの候補について，おのおの「が逮捕する」，「を逮捕する」，「が多発する」の項構造解析履歴を持つので，類似度はそれぞれ 0.4590, 0.7568, 0.3777 と求まり，最も類似度の高いものを選べば，それだけで正しく項を同定できることが分かる．

4. 格構造と項構造解析履歴の類似度を用いた述語項構造解析実験

4.1 実験内容

ベースラインに

- **CHAIN_LENGTH**　項候補が項になった回数
- **USED**　項候補が以前の文で項になったかどうか
- **CASE_SIM**　格構造と項構造解析履歴の類似度（提案手法）

の素性をそれぞれ組み合わせて，性能の違いを比較する実験を行った．なお本実験では，項構造解析履歴は文内項のみ参照し，それらの解析と名詞句の照応解析はすべて正しく行われたと仮定した．

格構造の類似度計算には，河原らが，web から収集した約 5 億文に対して JUMAN[⋆1] と KNP[⋆2] を用いて形態素解析と構文解析したもの[9] より，述語と名詞の助詞を介した係り受けの対計 1,101,472,855 対[⋆3]を用いた．

4.2 訓練・評価データ

本実験の学習と評価には NAIST テキストコーパス 1.4β[10] を用いた．NAIST テキストコーパス 1.4β は京都大学テキストコーパス Version 3.0 [⋆4]を元に，1995 年 1 月 1 日から 17 日までの全記事（約 2 万文）と 1 月から 12 月までの社説記事（約 2 万文）の計約 4 万文に対して，述語の格関係，事態性名詞[⋆5]の格関係，名詞間の照応関係をアノテートしたコーパスである．

⋆1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
⋆2 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
⋆3 異なり総数はそれぞれ，動詞:約 801 万，名詞:約 288 万，対:約 15994 万である
⋆4 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html
⋆5 動詞から派生した名詞やサ変名詞などの，動作を表す名詞のことで，述語と同様に格関係が定められる．本研究では対象としない．

図 1　トーナメントモデルを用いた項同定
Fig. 1　Argument identification with Tournament model

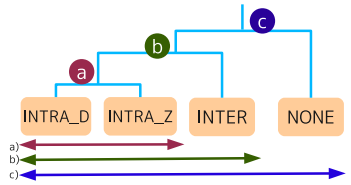
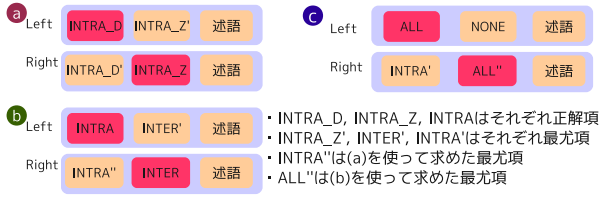


図 2　トーナメントモデルの学習
Fig. 2　Training of Tournament model



本実験では，NAIST テキストコーパス 1.4β から一部を除いた[⋆6]ものを対象に，記事ごとにランダムに並び替えた後，5 分割の交差検定を行った．また，対象とする文章に対して MeCab 0.98[⋆7] と CaboCha 0.60pre4[⋆8]を用い形態素解析・係り受け解析・固有表現解析を行った．

4.3 項同定モデル

本実験では，述語項構造解析を 2 つの段階に分けた．

1 段階目では，各述語に対して，項が仮に文内項（述語に直接係っているもの）・文内項（述語に直接係っていないもの）・文間項である場合，最も尤らしい項はどれであるかを，それぞれの項同定モデルを用いて求める．なお，本論文では以下それぞれの項を *INTRA_D, INTRA_Z, INTER* とよぶ．項を同定するモデルにはトーナメントモデル[5] を用いた．

2 段階目では，1 段階目で求めた INTRA_D, INTRA_Z, INTER の最尤項に対してどれが着目する述語（以下単に述語という）の項であるのか，もしくは述語が項を持たないのかを判定するために，トーナメントモデルを用いた．2 段階目の判定は**図 1** のように

(a) INTRA_D, INTRA_Z のどちらが述語の項らしいか
(b) (a) で勝ち上がってきたものと INTER のどちらが述語の項らしいか
(c) (b) で勝ち上がってきたものが述語の項であるかどうか

の 3 つの 2 値分類モデルに分割して行う．

学習は (a),(b),(c) の順で行い，事例の作成は**図 2** のように行う．

⋆6 タグの誤りのため 11 件の記事を除いた
⋆7 http://mecab.sourceforge.net/
⋆8 http://chasen.org/~taku/software/cabocha/

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

4.4 素性と学習器

ベースラインの素性として，

- 述語の語彙・統語情報に関する素性
- 先行詞候補に関する語彙・統語・意味情報，出現位置に関する素性
- SRL に関する局所文脈に関する素性
- 述語と項候補の共起頻度等の意味に関する素性

といった，飯田らの素性[11] を用いた．

各モデルの分類器には Support Vector Machine (SVM)[12] を用いた．[⋆1] カーネルは線形カーネルを用い，パラメータはデフォルト値とした．SVM は高い汎化能力を持ち，高次元の素性集合を用いても過学習しにくいとされ，形態素解析や係り受け解析などで使われている．[13][14]

4.5 評価指標

本実験では 2 つの性能変化を調べることで，提案手法の性能を評価する．

1 つ目は文間項同定の性能変化である．まず，文間項をもつ各述語に対してそれぞれ信頼度付きで最尤項を求める．信頼度の算出には飯田らが[5] で提案したトーナメントモデルの信頼度を用いた．次に信頼度の高いものから順に並べ，しきい値 θ_{inter} を変化させることで Precision と Recall のトレードオフがどのように変化したのかを評価する．Precision, Recall を次式を用いて計算した．

$$Precision = \frac{\text{信頼度が}\theta_{inter}\text{以上の正しく同定できている文間項の数}}{\text{信頼度が}\theta_{inter}\text{以上の文間項の数}}$$

$$Recall = \frac{\text{信頼度が}\theta_{inter}\text{以上の正しく同定できている文間項の数}}{\text{テストに用いた文間項の数}}$$

2 つ目はシステム全体から見た文間述語項構造解析の性能変化で，INTRA_D, INTRA_Z, INTER について Precision，Recall，F-measure を次式を用いて計算した．

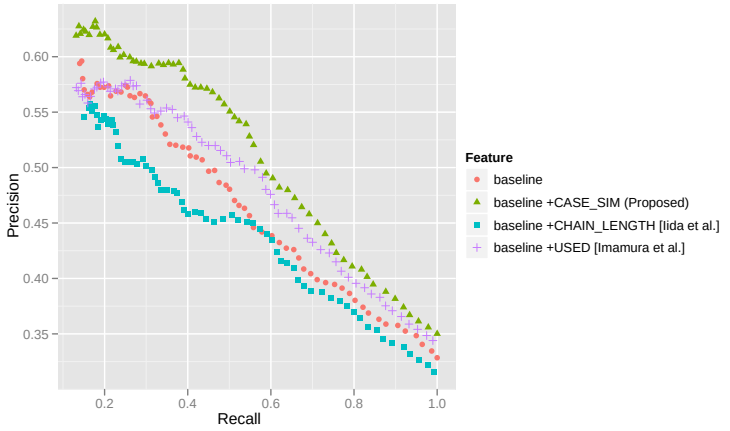
$$Precision = \frac{\text{システムの出力のうちの正解数}}{\text{システムが出力した項数}}$$

$$Recall = \frac{\text{システムの出力のうちの正解数}}{\text{テストに用いた項数}}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

⋆1 実装は LIBLINEAR(http://www.csie.ntu.edu.tw/~cjlin/liblinear/) を用いた．

図 3 ガ格文間述語項構造解析精度の Precision-Recall 曲線較
Fig. 3 Comparison of Precision-Recall curve for inter-sentential argument identification of nominative case



4.6 実験結果

文間項同定の Precision-Recall 曲線を**図 3** に示す．図より，提案手法はベースラインや CHAIN_LENGTH 素性・USED 素性と比較して，Precision, Recall ともに高く，項を同定できることがわかる．

次に，文間項同定に加え，文内の項同定も含めたシステム全体の評価結果を**表 9** に示す．提案手法を用いるとは，INTRA_Z では CHAIN_LENGTH 素性や USED 素性の方が Precision が高いが，INTRA_D の解析精度を犠牲にせずに INTER の Precision の向上に大きく寄与していることが分かる．

また，提案手法を CHAIN_LENGTH 素性や USED 素性と組み合わせて用いることで，さらに INTER の精度が向上できることがわかった．

5. 誤り事例分析

提案手法を用いても解析に失敗した事例を調べることで，どのような問題が残っているかについて分析する．

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

表 9 ガ格述語項構造解析精度の比較
Table 9 Comparison of predicate argument structure analysis of nominative case

	INTRA_D			INTRA_Z			INTER		
	P	R	F	P	R	F	P	R	F
ベースライン	78.9	90.5	84.3	44.7	59.6	51.1	17.6	16.2	16.9
+A （CHAIN_LENGTH 素性）	79.9	90.4	84.8	51.4	60.1	55.4	17.9	24.9	20.8
+B （USED 素性）	80.0	91.0	85.1	50.6	61.0	55.3	17.3	21.9	19.3
+C （CASE_SIM 素性）	79.1	90.6	84.4	45.4	60.3	51.8	19.7	18.4	19.0
+A+C	80.0	90.3	84.9	54.6	61.3	57.8	17.7	26.8	21.3
+B+C	79.9	90.9	85.1	51.7	62.0	56.4	17.8	23.0	20.1
+A+B+C	80.0	90.6	85.0	54.9	61.4	58.0	18.0	27.1	21.6

P, R,F はそれぞれ Precision, Recall,F-measure を示す．P,R の単位は%.

5.1 コピュラ

実験に使用した NAIST テキストコーパスでは「名詞+だ」についても述語としてタグづけされているが，それらは他の動詞と振る舞いが異なるため，同じ解析手法で解くのは困難だと思われる.「だ」や「である」は，それ自体には意味はなく，名詞と連接して述語を形成するはたらきがある．このようなものをコピュラとよぶ．具体的には次のような事例があった.

この白書では、東京圏の マンション価格$_{ガ}$ が、サラリーマンの平均年収の五・六倍に当たる平均四千七百七十四万円にまで下がってきたことを明らかにした。
ピークでは八・五倍、前年は**五・八倍**であった。

優勝$_{ガ}$ はフジタ工業時代の第五十九回大会以来、１５年ぶり。
古前田監督はまだ選手としてプレーしていた。
もちろん、二十代のいまの選手たちにとっては「**昔話**」に違いない。

これらは単なる動詞と区別して，解析モデルを作ることで対処可能である.

5.2 機能動詞結合

今回の実験では,「逮捕する」「感激する」などのサ変動詞は 1 つの述語として扱ったが,「逮捕をする」といった「名詞+格助詞+する」の場合の述語は単に「する」として扱った．実際には述語「する」には内容的な意味はなく，その前の名詞が主な意味を持つ．このようなものは「機能動詞結合」とよばれる[15]．したがって，機能動詞結合において項構造を付与する場合には，その前の名詞を含めて扱うべきである，なお，他の機能動詞結合の例として「影響を与える」の「与える」などが挙げられる.

具体的な次のような事例があった.

結婚の時は仲人も**して**いただいた。

経済事犯であると同時に、議員の肩書を背景に**して**おり、汚職的な側面も持つ。

これらは機能動詞結合全体で 1 つの述語として扱った方が良いと考えられる.

5.3 述語の曖昧性

同じ述語でも格フレームによって異なる項分布をとることがある.

例えば,

首相の諮問機関である「経済審議会$_{ガ}$」が二十九日、「構造改革のための経済社会計画―活力ある経済・安心できるくらし―」と題する新計画を村山内閣に答申した。
（中略）
このとき、私たちは「数字のない計画は意味がない」と指摘、数字を早急に**詰める**必要性を強調した。

という文章では述語「詰める」が出てくる．ところが,「詰める」には少なくとも「瓶にジャムを詰める」といった用法と「話を詰める」といった 2 つの用法がある．提案手法ではこれらを区別していない.

また,

逮捕容疑$_{ガ}$ となった背任をはじめとして、業務上横領、偽証、さらには詐欺までも訴追された。
経済事犯であると同時に、議員の肩書を背景にしており、汚職的な側面も**持つ**。

という文章における「持つ」も同様で,「荷物を持つ」という用法と「～な要素を持つ」という用法があるが，これらも区別できない.

今回は格フレームを考慮せずに格構造の類似度を求めたが，動詞を格フレームごとに区別して類似度を求めることで，精度はさらに改善できると考える.

Vol.2011-NL-201 No.10
Vol.2011-SLP-86 No.10
2011/5/17

6. まとめと今後の課題

本論文では文間述語項構造解析において文脈情報を用いるための手法について論じた．

先行研究では文脈情報を述語項構造解析において扱うための方法として，センタリング理論に基づく方法と，項候補が項となった回数を情報として用いる方法が，提案されてきた．ところが，いずれの手法を用いても,「X を逮捕した」という文だけを手がかりに「自首した」のガ格項が X であると判定することはできなかった．そこで本論文では,「格助詞+動詞」を格構造と定義し，格構造の類似度と述語項構造解析の履歴を用いる手法を提案した．これにより,「X」が「逮捕した」のヲ格である場合,「自首した」のガ格が X である可能性が高いということを捉えられるようになった．

また，比較実験より，提案手法を用いることで，従来の文間述語項構造解析より精度が向上することが分かった．さらに誤り事例の分析より，提案する格構造の類似度のとり方では，

- 機能動詞結合といった動詞自体には意味を持たない場合，動詞間の類似度を捉えることができない
- 複数の意味をもつ動詞では類似度をうまく測れないことがある

といったことが分かった．

今後の課題は主に 3 つあり，

- コピュラを単なる動詞と区別して処理すること
- 機能動詞結合がおきている場合，それを 1 つの動詞として扱うこと
- 複数の語義を持つ動詞を区別して類似度をとること

である．現在の NAIST テキストコーパスでは,「詰める」や「持つ」など複数の格フレームを持つものでも特に区別をしていので，それらの学習やテストを行うために，述語の格フレームを付与することも今後の課題である．

参 考 文 献

1) Grosz, B.: Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics*, Vol.21, No.2, pp.203–225 (1995).
2) Nariyama, S.: Grammar for ellipsis resolution in Japanese, *Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation*, pp.135–145 (2002).
3) Iida, R., Inui, K., Takamura, H. and Matsumoto, Y.: Incorporating contextual cues in trainable models for coreference resolution, *Proceedings of the 10th EACL Workshop on the Computational Treatment of Anaphora*, pp.23–30 (2003).
4) Niyu, G., Jhon, H. and Eugene, C.: A statistical approach to anaphora resolution, *Proceedings of the 6th Workshop on Very Large Corpora*, pp.161–170 (1998).
5) 飯田龍，乾健太郎，松本裕治：文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定，情報処理学会論文誌， Vol.45, No.3, pp.906–918 (2004).
6) 飯田龍：照応解析のための文脈的手がかりを考慮した機械学習モデル，奈良先端科学技術大学院大学修士論文 (2004).
7) Kenji, I., Saito, K. and Izumi, T.: Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution, *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pp.85–88 (2009).
8) Chambers, N. and Jurafsky, D.: Unsupervised learning of narrative event chains, *Proceedings of Annual Meeting of the Association for Computational Linguistics-08 with the Human Language Technology Conference* (2008).
9) Kawahara, D. and Kurohashi, S.: Case frame compilation from the web using high-performance computing, *In Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp.1344–1347 (2006).
10) 飯田龍， 小町守，井之上直也，乾健太郎，松本裕治：述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から，自然言語処理， Vol.17, No.2, pp.25–50 (2010).
11) Iida, R., Inui, K. and Matsumoto, Y.: Zero-anaphora resolution by learning rich syntactic pattern features, *ACM Transactions on Asian Language Information Processing*, Vol.6, No.4, pp.1:1–1:22 (2007).
12) Cortes, C. and Vapnik, V.: Support-vector networks, *Machine learning* (1995).
13) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proceedings of the Empirical Methods on Natural Language Processing*, Vol.2004, pp.89–96 (2004).
14) Kudo, T. and Matsumoto, Y.: Japanese dependency analysis using cascaded chunking, *Proceedings of the Conference on Computational Natural Language Learning*, pp.63–69 (2002).
15) 村木新次郎：日本語動詞の諸相，ひつじ書房 (1991).