複数の述語項構造の同時解析手法に関する調査

大内 啓樹　　進藤 裕之　　Kevin Duh　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{ouchi.hiroki.nt6, shindo, kevinduh, matsu}@is.naist.jp

1 はじめに

述語項構造解析は，述語と項の意味関係 (格関係) を同定するタスクであり，「誰が何を何にどうした」という意味的な関係を抽出することを目的とする．従来の述語項構造解析では，各述語に対して，他の述語項との相互関係を考慮せず，独立に格関係を決定するというアプローチが主流であった．しかし，述語間には意味的な関連があり，述語と項の格関係を決定する上で役立つことが期待される．例えば，次の文を考える．

(1) 花子に$_i$　殴られた$_{ガ:i}$　太郎は$_j$　近くの$_k$
病院に$_l$　入院した.$_{ガ:j}$

この例文において，「入院した.」のガ格 (主体) が「太郎」ならば，「殴った」(「殴られた」の能動態) のガ格は「花子」である可能性が高い．逆に，「入院した.」のガ格 (主体) が「花子」ならば，「殴った」のガ格は「太郎」である可能性が高い．このように，ある述語と格関係にある項の決定は，他の述語項の格関係決定に影響を及ぼす．これらを踏まえ，[6] は，複数の述語項の関係を考慮し，文内のすべての述語のガ格が付与される項を同時に推定している．しかし，彼らの手法は解析対象が文内にある項のみに限られている．

日本語では，ガ格の項の省略 (**ゼロ代名詞**) が頻繁に現れ，それが指し示す要素 (**先行詞**) が当該文より前の文に現れることも多い．ゼロ代名詞の先行詞を同定するタスクを**ゼロ照応解析**と呼び，述語項構造解析と関連しているため，同時に行われることも多い．例えば，次の文を考える．

(2) 彼女は$_i$　時間通りに　学校に　着いた.$_{ガ:i}$
しかし，　(ϕ_i)　宿題を　家に　忘れた.$_{ガ:i}$

(2) の例文において，「ϕ_i」がゼロ代名詞であり，「彼女は$_i$」がその先行詞となる．そして，「忘れた.」のガ格として，ゼロ代名詞の先行詞である「彼女は$_i$」を求めなければならない．

そこで，本稿では，文間にまたがるゼロ照応の解析に向けて，[6] の手法を利用し，当該文内の項だけでなく，それより前の文に含まれる項も解析対象とした場合の性能の調査を行う．

2 先行研究

日本語述語項構造解析研究で用いられてきたコーパスの一つに，NAIST テキストコーパス [2] がある．このコーパスは約 40,000 文の新聞記事・社説から構成され，述語項及び照応，共参照情報が付与されている．述語項構造の意味役割として，ガ格 (主格)，ヲ格 (対象格)，ニ格 (与格) の 3 種類が定義され，アノテーションされている．

NAIST テキストコーパスを用いた日本語述語項構造解析に関する代表的な先行研究として，[4] と [3] の二つが挙げられる．彼らは，格が付与される項を同定する際，述語との相対的な位置関係によって，各項を次の 3 つに分類し，解析結果を報告している．

係り受け有 (*INTRA_D*)： 述語と直接係り受け関係にある項.
文内ゼロ (*INTRA_Z*)： ゼロ代名詞の先行詞であり，述語と同一文内に現れる項.
文間ゼロ (*INTER*)： ゼロ代名詞の先行詞であり，述語と異なる文に現れる項.

項と述語間に係り受け関係がある場合 (*INTRA_D*)，解析が比較的容易である一方で，文内ゼロ (*INTRA_Z*)・文間ゼロ (*INTER*) を対象とした解析 (ゼロ照応解析) は困難であることが報告されている (表 1 を参照)．

[4] は，決定リストを利用した手法を提案している．Support Vector Machine の学習で各素性の重みを獲得し，それらをスコア順にソートしたものを決定リストとし，述語項の解析に適用した．彼らは，動詞・形容

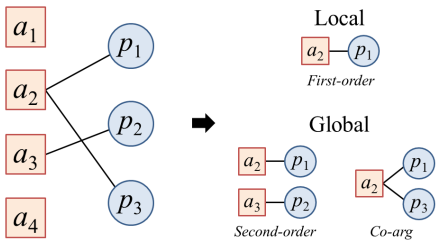

図 1　二部グラフモデル

詞などの一般的な述語だけでなく，事態性名詞についても項構造解析を行っている．

[3] は，最大エントロピーモデルを述語項構造解析に利用している．最大エントロピーモデルは，様々な素性を使用可能であることが特徴であり，彼らは大規模コーパスから獲得した言語モデルのスコアなどの文脈的な情報を利用し，高精度な解析結果を報告している．

彼らの手法は，各述語に対して，他の述語の項構造を考慮せず，項候補集合から最尤の項候補を選ぶことによって，格付与される項を独立に決定する．一方，[6] の手法は，複数の述語の項構造の相互関係を考慮し，文内におけるすべての述語に対して，項構造を同時に推定する．

3　複数の述語項構造の同時解析手法

3.1　二部グラフモデル

二部グラフモデル (図 1) は，複数の述語の項構造の相互関係を考慮することを可能にしている．[4] や [3] と同様に，格ごとに別々にモデルを構築する．ある格が付与される項を同定する際，その項候補集合 (図 1 中の a_i) と述語集合 (図 1 中の p_j) から二部グラフのノードを構成し，各述語ノードに項候補ノードを割り当てることで，各述語の項を表現する．具体的には，解析対象となる文 x に対して，二部グラフ $G(x) = (A_x, P_x, E_x)$ を構築する．A_x，P_x，E_x はそれぞれ，項候補集合，述語集合，エッジ集合を表す．項候補集合 A_x は項候補 (a_i) から構成され，ダミー項候補 NULL を含む．ダミー項候補は，述語が項をとらない場合や，文書中に項がない場合に割り当てられる．述語集合 P_x は，解析対象文 x 中のすべての述語 (p_j) から構成される．エッジ集合 E_x は，A_x と P_x 間のエッジ $(e_{a_i p_j})$ から構成され，エッジは各述語の格付与される項を表す．エッジは，各述語 p_j に対して，A_x に属するノード a_i から一本のみ張られる．

あるエッジ集合 E_x を持つ二部グラフ y に対して，重みベクトル θ と高次元素性ベクトル $\phi(x, y)$ の内積によってスコアを定義する．可能な二部グラフ集合 $G(x)$ から，スコア最大の二部グラフ $\hat{y}$ を，次式のスコア関数に従って求める．

$$\hat{y} = \underset{y \in G(x)}{\operatorname{argmax}}\ \theta \cdot \phi(x, y)$$

スコア関数内の重みベクトル θ は，機械学習手法を利用することで推定することができる．具体的には，正解データから二部グラフを構成し，$\hat{y}$ との誤差を少なくするよう学習する．本稿では，平均化パーセプトロン [1] を用いて，重みベクトル θ を推定した．

また，素性として，**局所的素性**と**大域的素性**を使用した [6]．局所的素性は，1 つの項候補と述語のペア (図 1 中の Local(*First-order*) 構造) に対して定義される素性である．一方，大域的素性は，複数の述語項間の相互関係 (図 1 中の Global(*Second-order*，*Co-arg*) 構造) に基づいて定義される素性である．これにより，各二部グラフに付与されるスコアは，複数の述語項の関係を考慮したものとなる．

しかし，大域的素性を用いることにより，スコア最大の二部グラフの探索は困難になる．その解決法として，**乱拓化山登り法** (Randomized Hill-Climbing) を利用し，近似的にスコア最大の二部グラフを探索する．

3.2　乱拓化山登り法

乱拓化山登り法は，依存構造解析において，依存構造木の探索のために提案された手法である [5]．本研究では，[6] が乱拓化山登り法に基づいて提案した，二部グラフ探索のための手法を用いる．アルゴリズムの概要は以下の通りである．

1. 初期二部グラフをランダムにサンプリングする．
2. 二部グラフのスコアが改善するよう，各述語ノードのエッジを一本ずつ替える．
3. どの一つの述語のエッジを替えても，それ以上スコアが向上しなくなるまで，2. を繰り返す．

このアルゴリズムを K 回繰り返し，得られた K 個の二部グラフの中で，スコア最大のグラフを最終的な解として選ぶ．この K 回の繰り返しを，**ランダムリスタート**と呼ぶ．ランダムリスタートの回数が多ければ，より良い局所解に辿り着く可能性が高くなり，解析精度向上が期待できる一方で，解析に時間を要するため，解析精度と時間のトレードオフが発生する．

3.3 文外の項候補の選出

本研究では，ガ格の文外項 (文間ゼロ) も解析対象とするため，解析対象の述語が含まれる文の一つ前の文に含まれる名詞句を項候補とする．これにより，正解となるガ格の文外項の 54% ほどをカバーすることが可能となる一方，項候補数が一文平均 5.4 増えるため，探索が困難になることが予想される．そのため，本研究の評価実験において，既存手法と解析精度を比較することにより，解析が有効に機能するかを明かにする．

4 評価実験

文外項も含めた解析に [6] の同時解析手法を適用し，その有効性を調査するため，評価実験を行った．解析対象をガ格の項 (*INTRA_D*，*INTRA_Z*，*INTER*) に設定した．また，評価・解析は，文節単位で行った．

4.1 データセット

データセットとして，述語項構造・共参照情報が付与されたコーパスである，NAIST テキストコーパス 1.5[2] を用いた．先行研究などで採用される標準的なデータ分割法を採用し，モデルの訓練・開発・評価を行った[*1][4].

実験で用いる品詞タグ，文節境界，係り受け情報は NAIST テキストコーパスに付与されているアノテーションを利用した．本実験では，京大格フレームやその他の外部資源は一切利用していない．

4.2 ベースライン

同時解析手法と比較するためのベースラインとなる手法として，[3] の手法を採用した．これは，各述語に対してそれぞれの項候補とのスコアを点推定し，スコア最大の項候補を選ぶことによって，格を付与する項を決定する．素性として，[3] で使用された素性のうち，Additional Features として定義された素性以外の基本素性を使用する．

4.3 実装詳細

解析の際に用いる項候補として，ベースライン・提案手法どちらにおいても，文内に含まれる全文節と，当該文の直前の文において主辞が名詞である文節とした．解析対象の述語は，NAIST テキストコーパスでアノテーションされているものを利用した．

同時解析手法におけるランダムリスタート回数は，開発データを使用し解析精度がほぼ収束した回数である 50 回 ($K = 50$) に設定した．解析結果として，独立に 10 回解析した結果の平均を報告する．モデルの学習アルゴリズムとして，ベースライン・同時解析手法どちらも，平均化パーセプトロンを採用し，イテレーション数は 15 に設定し訓練した．

[*1] 訓練には 1 月 1-11 日の記事と 1-8 月の社説，開発には 1 月 12，13 日の記事と 9 月の社説，評価には 1 月 14-17 日の記事と 10-12 月の社説を利用した．

5 結果と考察

表 1 は，評価データにおけるベースラインの手法と同時解析手法の解析結果を示している．また，実験設定などが異なるため，厳密に比較することは困難ではあるが，既存研究との比較のため，Taira et al. (2008) [4] と Imamura et al. (2009) [3] の結果も掲載する．

5.1 ベースラインとの比較

同時解析手法では，係り受け有 (*INTRA_D*) の述語項構造の解析結果が，F 値で 88.12% となり，ベースラインを約 2.1 ポイント上回った．また，文内ゼロ (*INTRA_Z*)，文間ゼロ (*INTER*) のそれぞれにおいても，同時解析手法の F 値は 35.27%(+1.24)，10.26%(+2.61) であり，ベースラインを上回った．この結果から，直接係り受けのある項やゼロ照応関係にある項の格関係の決定に，同時解析手法が有効に機能していることがわかる．

5.2 既存研究との比較

既存研究と比較すると，直接係り受けの有る項の解析結果においては，同時解析手法が Imamura et al. (2009) を約 1 ポイント，Taira et al. (2008) を約 13 ポイント上回った．文内ゼロに関しては，Imamura et al. (2009) が 50.0% であり，35.27% である同時解析手法を約 15 ポイント上回っているが，Taira et al. (2008) は 30.15% であり，同時解析手法が約 5 ポイント上回っている．文間ゼロに関しては，Taira et al. (2008) が最も高い精度 (23.45%) を記録しており，同時解析手法と比べ約 13 ポイント，Imamura et al. (2009) と比べ約 10 ポイント上回っている．

これらの結果を見ると，同時解析手法と Imamura et al. (2009) の手法の解析結果において，直接係り受けが有る項の解析精度が高く (ともに 80% 後半)，文間ゼロが低い (ともに 10% 前半) という傾向が見られる．一方，Taira et al. (2008) の手法は他の手法と比べ，直接係り受けの有る項の解析精度は高くない (75.53%) が，文間ゼロの解析精度は同時解析手法や Imamura et al. (2009) の手法と比べ，10 ポイントほど高くなっている．

	INTRA_D			*INTRA_Z*			*INTER*		
Model	P	R	F_1	P	R	F_1	P	R	F_1
ベースライン	85.07	86.92	85.99	29.27	40.64	34.03	18.50	4.82	7.65
同時解析手法	87.79	88.46	**88.12**	29.92	42.96	35.27	25.16	6.44	10.26
Taira et al. (2008) [4]	-	-	75.53	-	-	30.15	-	-	**23.45**
Imamura et al. (2009) [3]	85.2	88.8	87.0	58.8	43.4	**50.0**	47.5	7.6	13.1

表 1　評価データを用いた解析結果．*INTRA_D*=係り受け有．*INTRA_Z*=文内ゼロ．*INTER*=文間ゼロ．P=適合率．R=再現率．F_1=F 値.

これは，同時解析手法の局所的素性として，Imamura et al. (2009) の使用した素性と同様の素性[*2]を利用しており，手法も類似した点[*3]があるため，二つの手法の解析結果は同じような傾向を示したと考えられる．このことから，どのような素性や手法がどの種類の項に効果的か今後詳しく調査をしたい．

また，文内・文間ゼロの解析において，同時解析手法が高精度を達成できていない要因として，外部資源を用いていないため，語の選好性などの情報が不十分であることが挙げられる．ゼロ照応の場合，構文的手がかりが有効に働かない場合が多いため，述語と項の共起情報やその他の意味的な情報が重要となる．今後，格フレームなどの外部資源を利用し，そのような情報を素性として組み込み，改善を試みたい．さらに，文間ゼロの解析精度向上のためには，当該文より前の文から適切に項候補を選出する必要があると思われる．本研究のベースラインの手法や同時解析手法では，当該文の直前の文の名詞句を項候補としているため，より効率的な項候補の絞り込みの方法が今後の課題となる．

6　おわりに

本稿では，文間ゼロ照応解析に向けた複数の述語項構造の同時解析手法に関して調査した．評価実験を通して，同時解析手法は直接係り受けがある項の格関係同定において高い精度を示した．その理由として，文内の複数の述語項構造の情報が，項構造を決定する際に相互に有用な手がかりとして働いていることが考えられる．一方で，ゼロ照応関係にある項の解析に関しては，改善の余地がある結果となった．これは，ゼロ照応関係にある項と述語は統語的な手がかりが乏しいため，語彙の共起情報や意味的な情報に頼ることになるが，外部資源を利用していない本研究ではその情報が不十分であった可能性がある．今後，格フレームなどの外部資源を利用した素性や，文外の項候補の有効な選出法などを調査したい．

参考文献

[1] Michael Collins. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*, pp. 1–8, 2002.

[2] Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 132–139, 2007.

[3] Kenji Imamura, Kuniko Saito, and Tomoko Izumi. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proceedings of ACL-IJCNLP*, pp. 85–88, 2009.

[4] Hirotoshi Taira, Sanae Fujita, and Masaaki Nagata. A japanese predicate argument structure analysis using decision lists. In *Proceedings of EMNLP*, pp. 523–532, 2008.

[5] Yuan Zhang, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Greed is good if randomized: New inference for dependency parsing. In *Proceedings of EMNLP*, pp. 1013–1024, 2014.

[6] 大内啓樹, 進藤裕之, Kevin Duh, 松本裕治. 複数の述語項関係を利用したゼロ照応解析. 情報処理学会自然言語処理研究会, NL-220, 2015.

[*2] ただし，Additional Features 以外.

[*3] *First-order* 構造の部分.