

識別的スクリプトモデルを用いた文間ゼロ照応解析

浜田 伸一郎

東芝ソリューション (株)

hamada.shinichiro@toshiba-sol.co.jp

1 はじめに

ある言語表現が、テキスト内で先行する表現と同じ内容や実体を指す現象を照応と呼ぶ。照応は節間や文間で発生し、文脈を形成する役割を果たす。このため照応解析は、要約・翻訳・検索・情報抽出など文脈理解を必要とする様々な応用技術において不可欠な基礎技術である。日本語では、後続の語を省略して照応関係を示す方法を取ることが多く、これを特にゼロ照応と呼ぶ。本稿では、ゼロ照応の中でも特に難しいとされる文間ゼロ照応解析について論じる。

図1は文間ゼロ照応の例である。文aおよび文bで省略されているガ格の照応先候補として「A社」と「B社」が挙げられるが、どちらかといえば文aの正解は「A社」、文bの正解は「B社」と思われる。この判断をするには、文法的手がかりだけでは不十分であり、ジャンクのスクリプト[13]のような手続き知識が必要であると考えられる。

手続き知識は、文間ゼロ照応に用途が限定されない広く汎用的な知識であるため、文間ゼロ照応モデルとは独立してモデル化すべきであると考えられる。そこで本稿では、手続き知識の計算モデル(これをスクリプトモデルと呼ぶ)の構築方法、およびそれを用いた文間ゼロ照応解析手法を提案する。

2 アプローチ

本稿が想定するスクリプトモデルは、述語項構造を手続きステップとみなした系列モデルである。ゼロ照応解析の先行研究においても、これに類するスクリプトモデルを構築し、そのモデルを用いて評価したスクリプトの適合性を解析器で利用するアプローチが様々な試みられてきた。表1は先行研究において用いられたスクリプト相当の計算モデルの一覧である。

モデルの多くは基本的に長さ1~2と短く¹、またいずれの手法も順序を扱わない。この背景には、スパースネス問題を回避する意図があるほか、元々関連述語対を抽出する研究[10, 14]において提案された手法であるという出自上の理由もあると思われる。ただしchamberモデル²は、順序は扱わないが任意長に対応している。

ところで長さ2以上の系列モデルを構築する場合、相関性の高い述語項構造群だけをフィルタする前処理が必須である。文章中の述語項構造の出現順序は非常に乱雑であり、単純な集計では傾向を捉えるのが難しいためである。フィルタ処理として最もよく使われる手法の1つにアンカー方式[10, 14]がある。表1でも長さ2以上のモデルの多くはアンカー方式を用いている。アンカーとは文脈を生み出す事物のことであり、アンカー方式とは、文章中に現れた各事物ごとに、それに関して発生した述語+表層格スロット

A社はB社に対し業務上の過失を認めた。

a: (φが) 再発防止を誓った。

b: (φが) 賠償請求することにした。

図1: 文間ゼロ照応の例

(以降イベントスロットと呼ぶ)の系列を統計処理対象とする方式である。イベントスロット同士が項を共有することになるので、項共有モデルとも呼ばれる。

ところで、以上に紹介したモデルはいずれも、スクリプトの傾向を十分捉えるには問題があると考えられる。

まず長さについてであるが、人手で作成されたスクリプト資源であるFramenet[1]に含まれるフレームの述語数は平均6程度との報告があるなど[2]、スクリプトは比較的長い系列の中で認識される傾向にある。図2は[2]で挙げられた長さ5のスクリプトの例である。

また順序関係も重要であり、たとえば図2のスクリプトをランダムシャッフルすると、違和感のあるスクリプトになることが多い。しかし部分的なスキップ・割込・順序転倒などの変形はむしろ起こりやすく、たとえば、構文構造により文書内でのステップの出現順序が転倒する可能性があるほか、図2のスクリプト例では、いずれかのステップが脱落したり、途中に「後悔する:ガ格」が挿入されても違和感はない。そこで本稿では、長い系列に対する順序制約と部分的なスキップ・割込・順序転倒に対する許容性とを併せ持つ系列モデルの実現を目標とする。

またsasanoモデル以外いずれのモデルも、述語語義の曖昧性を除去せずに、述語項構造を手続きステップに対応付けている。しかし述語項構造解析の多くの研究[9, 16]では、述語と項の間を取り持つ中間変数として述語語義を用いており、述語と項の関係は述語の語義・用法毎に分けて説明するのが有効であると認識されている。本稿でもこの認識に従い、語義曖昧性を除去した述語を用いてスクリプトモデルを構築する。

項区別は、たとえば図2に示されるスクリプトは項が「犯人」だと適合的だが「ハンブティ・ダンブティ」だと不適合的であるというように、アンカーに応じたスクリプトの適合性を評価する仕組みである³。本稿でもこの仕組みを採用する。ただしスパースネス対策のため、意味クラス・固有表現タイプ・表層などから適切な粒度が自動選択されるよう、3.2節で述べる統合的な統計の枠組みの中に組み込む。

最後に複数項についてだが、笹野はガ格/ヲ格/ニ格の同時生起をモデル化している。系列モデルではスパースになりすぎるため、本稿では扱わない。

3 スクリプトモデル

スクリプトモデルの構築方法について述べる。

¹長さ1のモデルは系列をなしていないため、述語項構造単体の適合性を評価している。

²この手法は、厳密にはゼロ照応解析に用いられた手法ではないが、gerberモデルのアイデアの元となっている。

³[2]では項区別のあるスクリプトモデルのことをタイプ付き談話鎖と呼んでいる

表 1: 既存のゼロ照応解析で用いられている各種スクリプトモデル

モデル名	長さ	順序	項区別	複数項	述語語義	フィルタ処理	適合性指標	応用先
imamura [7]	1	—	○	×	×	—	生起確率	日本語ゼロ照応
sasano [12]	1	—	×	○	○	—	生起確率	日本語ゼロ照応
iida [18]	2	×	×	×	×	アンカー	共起確率ベース	日本語ゼロ照応
gerber [3]	2	×	×	×	×	アンカー	PMI	英語事態性名詞ゼロ照応
hayashibe [4]	2	×	×	×	×	項分布 ^a	Jensen-Shannon 距離	日本語ゼロ照応
chamber [2]	任意	×	○	×	×	アンカー	PMI	—

^a似たような項分布を持つ述語は項共有しやすいと仮定する方式

- (1) 捜索する:ヲ格
- (2) 逮捕する:ヲ格
- (3) 抗弁する:ガ格
- (4) 有罪宣告する:ヲ格
- (5) 処刑する:ヲ格

図 2: アンカー方式に基づくスクリプトの例

- (1) 自衛隊の出動 → 自衛隊が出動する
- (2) 日米交渉 → 日本と米国が交渉する

図 3: 事態名詞の動詞を用いた言い換え

[中国] 発展:動32:ga 拡大:動1:ga 成長:動1:ga
 [総会] 失敗:動41:ga 経験:動11:ga
 [党指導者] 重視:動1:ga 節約:動1:ga 持続:動1:ga 切り替える:動3:ga
 [環境面] 条件:判1:ga 重視:動1:wo

図 4: イベントスロット系列の収集例

積む/つむ:動1

- [ガ格] 選手:17, 自分:14, 人:10, 全員:7, ..
- [ヲ格] 経験:37342, 体験:1363, ...

積む/つむ:動3

- [ヲ格] 荷物:3592, 石:1074, 荷:884, 煉瓦:836, ..
- [ニ格] 車:739, トラック:100, 船:77, ..

図 5: 京都格フレームの抜粋

3.1 イベントスロット系列の収集

NAIST コーパス 1.5⁴[6] を対象として、イベントスロット系列の収集処理を行う。具体的には、ゼロ照応タグを用いて述語の省略項をすべて埋めた後、共参照タグを用いて同定されたアンカーごとにイベントスロット系列を収集する。

収集対象とする述語の品詞は、動詞・形容詞・コピュラ・名詞の4種類とする。ただし動詞については可能な限り能動態に変換し、名詞については図3に示すように動詞を用いた同じ意味の表現に機械的に置き換える。これにより表現の多様性を抑制しスパースネスを低減する。

収集対象とする表層格タイプは、人手によるタグが付与されているガ格・ヲ格・ニ格の他、述語項構造解析器である knp3.01[9] を用いて機械的にタグ付けしたデ格・ト格・カラ格・ヘ格・ヨリ格・マデ格・修飾格・時間格・トイウ格・無格・ノ格・ガ2格・外の関係を加えた16種類とする。

収集対象とする述語には、knp を用いて語義曖昧性を除去したもの(具体的には述語基本形に語義番号などが付与されたID)を用いる。ただし knp が出力する語義は、クラスタリングによって機械的に獲得した語義モデルに基づいているため、必ずしも人間にとって自然な分け方にはなっておらず、過分割気味である。また knp は、語義同定に失敗した場合、匿名語義を割り当てる⁵。

入力文章「...サラリーマンの家はどんどん遠くなった。遠距離通勤は、肉体的疲労のみでなく...」の場合、アンカー「サラリーマン」に対し、事態性名詞「通勤」を動詞化して得た「～が通勤する」、事態性名詞「疲労」を動詞化して得た「～が疲労する」、の2つの手続きステップが存在し、knp による語義曖昧性除去処理を通じて、最終的にイベントスロット系列「[サラリーマン] 通勤:動1:ga 疲労:動1:ga」を得る。図4はこのようにして収集したイベントスロット系列群の例である。

また第2の学習データとして京大格フレームコーパス1.0[8] (以降kcfと呼ぶ)を用いる。これはknpの述語語義モデルの訓練のために収集した統計データを一般公開したものであり、16億文のWebテキストでの述語語義・表層格タイプ

(イベントスロット)ごとの項の出現頻度が含まれている(図5)。これをunigramモデルの構築に利用する。

3.2 識別モデル

続いて、収集したイベントスロット系列群を用いて、スクリプトの確率モデルを構築する。2節で述べたように、長い系列における順序制約、部分的なスキップ・割込・順序転倒、項区別などの複合的な性質を備えたスクリプトモデルを実現したい。これに対し、言語モデルなどでよく用いられる生成モデルは、基本的には単一種類の統計量しか扱えない。そこで本稿では、複合的な性質を統合的に扱うことができる最大エントロピー法[11]を用いる。最大エントロピー法は、先ほど述べた系列の各種性質を識別モデルの素性として扱い、それら素性群をすべて含む最適化問題を通じて素性群の重要度を決定し、文脈を与えたときの後続イベントスロットの推定確率を出力するモデルを得る。本稿が用いる素性の種類とその発火例を表2に示す。

[1gram-kcf] kcfに含まれるイベントスロットごとの項の出現頻度に基づくunigram対数尤度を値に取る素性グループ。16億文のWebテキストでの頻度に基づくため信頼度が高い。ngramの平滑化の補助として機能する。

[ngram] 通常Ngramに関する素性グループ。ただし指定長以下のすべての通常Ngramを素性として生成する。これにより平滑化効果を得る。

[skipgram][11] 履歴の一部に任意のイベントスロットにマッチするワイルドカードを含むNgramに関する素性グループ。SkipGramと呼ばれる。ある程度順序を保ちつつ、部分的なスキップ・割込み・順序転倒をある程度許容する。また長距離の履歴を扱ってもスパースになりにくい。通常Ngramと同様、平滑化のため指定長以下のすべてのSkipGramを生成する。

[bow][11] 履歴に指定のイベントスロットが含まれるかを示す素性。bag-of-wordsまたはトリガーと呼ばれる。部分的なスキップ・割込み・順序転倒を良く許容するが順序を

⁴新聞記事に形態素・述語項構造・照応・共参照などのタグを付与したコーパス

⁵NAIST コーパス内の述語の語義曖昧性除去処理において、語義同定に失敗した割合は約12.6%だった

表 2: イベントスロット系列「[A1] S1 S2」に対する素性抽出例

素性グループ	発火する素性
lgram-kcf	S2(素性値はunigram対数尤度)
ngram	S1_S2 S2
skipgram	S1_* *_S2 *_* **
bow	S1 S2
arg&len	A1&*_* A1&*
arg&all	A1と上記各素性との組合せ

A1はアンカー, S1およびS2はイベントスロット, *はワイルドカード, &は同時発火を表す。lgram-kcfは実数素性, それ以外は2値素性。

扱えない。また長距離の履歴を扱ってもスパースにならない。

[arg&len] アンカーを示す素性と, ワイルドカードだけで構成されるskipgram(=系列長のみを示す)との組合せ素性。アンカーごとの系列長の傾向を捉えるため, 未知のイベントスロット系列などに対する推定の補助として機能する。表層・意味クラス・固有表現タイプなど, 異なる観点で表現した複数のアンカー素性群を用意する。これにより最適化処理の中で, 適切な粒度のアンカー素性が選択される。

[arg&all] アンカーを示す素性と, 他のすべての素性とを組み合わせた素性。同じイベント系列に対しアンカーに応じた尤度を推定できる。すなわち2節で述べた「項区別」の効果である。

以上のスクリプトモデルを, 最大エントロピー法に単純に実装すると, メモリに乗り切らず処理不可能である。なぜならパラメータ数はクラス数×素性数だが, クラス数はイベントスロット種類数に一致し約4.5万, 本節で説明した素性数は約90万(5節のシステムs369aaの場合), という規模だからである。そこでハッシュカーネル[17]を用いる。ハッシュカーネルは, スパースな素性ベクトルに対し, 内積値を良く保ちつつ次元圧縮する手法であり, ベクトル x の変換結果は次のような $\phi(x) \in R^t$ で表される。ただし, ξ, h はそれぞれ入力 x を $\{-1, +1\}$, $[1, \dots, t]$ に写像するようなハッシュ関数である。なお今回, 圧縮後の素性数 t は5000とした。

$$\phi_i(x) = \sum_{j: h(j)=i} \xi(j)x_j$$

4 文間ゼロ照応解析

続いて, スクリプトモデルを用いた文間ゼロ照応解析器の仕組みについて説明する。

文間ゼロ照応解析とは, 文間の距離にあるゼロ照応詞と先行詞のペアを同定するタスクである。ただし前提条件として, 形態素解析・係受解析・共参照解析・意味クラス解析・述語項構造解析・文内ゼロ照応解析が実施済みであり, 単語分割・品詞・係受・共参照・意味クラス・述語・述語語義・述語の非省略項および文内ゼロ照応項・未充足の格スロット・前方文脈での文間ゼロ照応項が分かっているものとする。以上の設定は, 述語項構造解析・文内ゼロ照応解析が実施済みという条件以外, 従来研究の多くと同じ設定である。

本稿では, 述語項構造解析および文内ゼロ照応解析によって充足されなかった格スロット群をゼロ照応詞の候補とすることを推論の出発点とし, 学習モデルには, 棄却付きランキング学習[5]を用いる。この手法は, 未充足格スロットと文間距離にある各先行名詞候補とのペアを事例として扱い, まずランキング学習によって最尤の先行詞候補を選択する。ただし未充足格スロットはゼロ照応詞とは限らないため, 先行詞が存在しない可能性がある。そこで引き続き, 得られた最尤の先行詞候補と未充足格スロットのペア事例を棄却するかどうかの2値分類を行う。以上の手法には, 述語

表 3: 文間ゼロ照応解析の素性

先行詞	文法意味構造	表層; 品詞; 指示詞タイプ; 助詞の種類 意味クラス; 主体か 主節へ係るか; 連体節か; 文頭か; 先頭文か
照応詞 (述語)	文法構造	述語の品詞; 態 連体節か
ペア	述語項構造 文脈	項は主体か; 格の種類; 間接参照か; 距離 SRLスコア; スクリプトスコア

の照応性に関する2値分類を先に行ってから先行詞をランキングする手法と比べ, 多くの手がかりを2値分類で利用出来るという利点がある。

表3は学習処理で用いる素性の一覧である。既存研究で提案された様々な特徴群[5, 8, 15, 7]に, スクリプトスコア素性を加えた構成となっている。このスクリプトスコア素性には, 対象事例に関するスクリプトとしての尤度を各種スクリプトモデルを用いて算出した結果が入る。たとえば, 対象事例のイベントスロットを S_y とし先行詞候補を A_1 とするとき, A_1 をアンカーとした先行文脈内のイベントスロット系列「[A1]S1S2...Sn」に S_y を加えた系列についてスクリプトモデルを用いて評価して得た尤度が素性値となる。

5 評価実験

ゼロ照応解析における各種スクリプトモデルの有効性を確認する実験を行った。実験データにはNAISTコーパス1.5を用い, [15, 7, 4]と同じ割り当て⁶とした。解析対象とする述語の品詞は, 動詞・形容詞・コピュラのみとし, 名詞は扱わない。ゼロ照応詞候補の格種別は, ガ格・ヲ格・ニ格の3種類とする。また扱う先行詞候補数が膨大になるのを防ぐため, ゼロ照応詞候補と先行詞候補との距離上限を形態素単位で190とした。先行詞の95%がこの距離に収まっている。これは再現率の上限となる。

評価結果を表4に示す(比較対象のシステムおよび評価方法の詳細については, 表4の脚注を参照せよ)。

照応ペア抽出タスクにおいて, いずれの提案システムも, 既存の短距離モデルに基づくシステムであるsasano2(F値:23.21%)・gerber2(F値:28.53%)に対する性能改善が見られた。このことは長い系列モデルの有効性を示していると考えられる。提案システム間での性能差はあまり大きくないが, s300に対しs335の性能はやや向上しており(F値:29.24%→30.88%), SkipGramやトリガーの有効性を示していると考えられる。一方, s335に対しs369の性能はむしろ低下していることから(F値:30.88%→29.24%), SkipGramやトリガーの長さの適切な設定は, 長さ3~5付近であると考えられる。1節で述べたようにFramenetに含まれるフレームの述語数は平均6程度との報告があるが, それより短い系列が最適設定となるのは, 学習データのスパース性が要因と思われる。またアンカーを扱うシステムs369a, s369aaは, s369より高い性能を示したことから, アンカーはスクリプトの推定に役立つと考えられる。特に全システム中で最良性能を示したs369aが用いる「アンカーに基づくスクリプト長の推定」は補助として有効と考えられる。

一方, 先行詞同定タスクについては, 既存システムを含めた全システムにおいてほとんど性能差が見られなかった。したがって照応ペア抽出タスクでのシステム間の性能差は, ランキング処理後の棄却処理によるものということになる。

⁶訓練データ:1月~8月の社説記事+1/1~1/11の一般記事, 開発データ:9月の社説記事+1/12~1/13の一般記事, 評価データ:10月~12月の社説記事+1/14~1/17の一般記事

表 4: 各種スクリプトモデルを用いた文間ゼロ照応解析の性能比較

	スクリプトモデル							照応ペア抽出			先行詞同定	
	kcf	ngrm	skip	bow	a&l	a&a	MRR	P	R	F ₁	P	
s300	y	3	0	0	n	n	0.1144	44.97	21.66	29.24	47.04	
s335	y	3	3	5	n	n	0.1152	46.25	23.17	30.88	47.00	
s369	y	3	6	9	n	n	0.1146	46.78	21.66	29.24	47.30	
s369a	y	3	6	9	y	n	0.1130	47.79	22.91	30.98	47.08	
s369aa	y	3	6	9	y	y	0.1121	47.91	21.62	29.70	47.12	
sasano2	[12]の再現版, 長さ1, 複数項							—	44.91	15.65	23.21	47.00
gerber2	[3]の再現版, 長さ2, PMI							—	46.71	20.54	28.53	47.38
taira	[15]の結果(転載), スクリプトモデルなし							—	—	—	23.45	—
imamura	[7]の結果(転載), 長さ1							—	47.5	7.6	13.1	—
hayashibe	[4]の結果(転載), 長さ2, JS距離							—	24.86	21.57	23.08	—

s300～s369aaは提案した各種スクリプトモデルを組み込んだゼロ照応解析システム。kcf～a&aはそれぞれ表2の素性グループを示し、y/nは採否、数字は回数(履歴長+1)を示す。sasano2とgerber2は、各文献のスクリプトモデルを再現し、本稿ゼロ照応システムに組み込んだシステム。taira, imamura, hayashibeは各文献の結果を転載したもの。ただし本稿とは実験条件が述語項構造解析と文内ゼロ照応解析も同時に解く点で異なる。

MRRは $1/|N| \cdot \sum_{n \in N} 1/\text{rank}(n)$ で表されるランキング性能の指標。ただしNは事例数、rank(n)は事例nに出現する先行詞候補をランキングした時の先行詞の順位を表す。Pは精度、Rは再現率、F₁はF値($\beta=1$)を示す。照応ペア抽出タスクは、ゼロ照応詞(イベントスロット)と先行詞のペアを同定するタスクである。先行詞同定タスクは、与えられた照応詞に対する先行詞を同定するタスクであり、照応ペア抽出タスクの一部をなすランキング処理に対応する。ただしこのタスクでは、先行詞は必ず存在するという設定とし、そのため評価指標は精度のみとする。

このことは、スクリプト同士でどれが尤もらしいか判定するのは困難だが、スクリプトが完了しているか判定するのは比較的容易であることが背景にあると思われる。

最後にMRRはスクリプトモデル単独でのランキング性能に関する指標である。しかし目立った性能差は数字としては現れなかった。前述のように、候補の中からどれが尤もらしいかを判定するのは困難であるため、数字に表れるほどの顕著な差が生じなかったのではないと思われる。

6 おわりに

本稿では、スクリプトモデルを用いた文間ゼロ照応解析を提案した。このスクリプトモデルは、識別モデルに基づいており、長い系列に対する順序制約と部分的なスキップ・割込・順序転倒・項区別などの複合的性質を備えた確率モデルとなっている。文間ゼロ照応解析への適用評価実験を行ったところ、短い系列のスクリプトモデルを用いる既存手法に対する性能改善が確認できた。

参考文献

- [1] C.F. Baker, C.J. Fillmore, and J.B. Lowe. The berkeley framenet project. In *ACL-COLING*, pp. 86–90, 1998.
- [2] N. Chambers and D. Jurafsky. Unsupervised learning of narrative schemas and their participants. In *ACL*, pp. 602–610, 2009.
- [3] M. Gerber and J.Y. Chai. Beyond nombank: A study of implicit arguments for nominal predicates. In *ACL*, pp. 1583–1592, 2010.
- [4] Yuta Hayashibe, Mamoru Komachi, and Yuji Matsumoto. Japanese predicate argument structure analysis exploiting argument position and type. In *IJCNLP*, pp. 201–209, 2011.
- [5] R. Iida, K. Inui, and Y. Matsumoto. Anaphora resolution by antecedent identification followed by anaphoricity determination. *TALIP*, Vol. 4, No. 4, pp. 417–434, 2005.
- [6] R. Iida, M. Komachi, K. Inui, and Y. Matsumoto. Annotating a japanese text corpus with predicate-argument and coreference relations. In *LAW*, pp. 132–139, 2007.
- [7] K. Imamura, K. Saito, and T. Izumi. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *ACL-IJCNLP Short Papers*, pp. 85–88, 2009.
- [8] D. Kawahara and S. Kurohashi. Case frame compilation from the web using high-performance computing. In *LREC*, pp. 1344–1347, 2006.
- [9] D. Kawahara and S. Kurohashi. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *NAACL-HLT*, pp. 176–183, 2006.
- [10] V. Pekar. Acquisition of verb entailment from text. In *NAACL-HLT*, pp. 49–56, 2006.
- [11] R. Rosenfeld. A maximum entropy approach to adaptive statistical language modelling. *Computer speech and language*, Vol. 10, No. 3, p. 187, 1996.
- [12] R. Sasano and S. Kurohashi. A discriminative approach to japanese zero anaphora resolution with large-scale lexicalized case frames. In *Proceedings of the ACL-IJCNLP 2011*, pp. 758–766, 2011.
- [13] R.C. Schank and R.P. Abelson. Scripts, plans, goals and understanding: An inquiry into human knowledge structures. 1977.
- [14] I. Szpektor and I. Dagan. Learning entailment rules for unary templates. In *COLING*, pp. 849–856, 2008.
- [15] H. Taira, S. Fujita, and M. Nagata. A japanese predicate argument structure analysis using decision lists. In *EMNLP*, pp. 523–532, 2008.
- [16] Y. Watanabe, M. Asahara, and Y. Matsumoto. A structured model for joint learning of argument roles and predicate senses. In *ACL*, 2010.
- [17] K.Q. Weinberger, A. Dasgupta, J. Langford, A. Smola, and J. Attenberg. Feature hashing for large scale multitask learning. In *ICML*, pp. 1113–1120, 2009.
- [18] 飯田, 徳永. 述語対の項共有情報を利用した文間ゼロ照応解析. 言語処理学会年次大会, 2010.