情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2010-NL-198 No.9
2010/9/17

小規模タグ付きコーパスと自動獲得した大規模語彙知識を用いた識別モデルに基づくゼロ照応解析

笹野遼平[†1]　黒橋禎夫[†1]

本稿では，比較的小規模の関係タグ付きコーパスから獲得した構文的な手掛かりと，大規模なタグなしコーパスから獲得した語彙的手掛かり利用した識別モデルに基づくゼロ照応解析モデルを提案する．提案モデルではゼロ照応を解く際，ゼロ代名詞とその先行詞を格ごとに個別に対応付けるのではなく，各用言に対し適切な格フレームを選択し，格フレームとテキスト中に出現した談話要素の対応付け全体で評価し，適切な対応付けを行う．

A Discriminative Approach to Zero Anaphora Resolution with Small Annotated Corpus and Lexicalized Case Frames

Ryohei Sasano[†1] and Sadao Kurohashi[†1]

This paper presents a discriminative, entity-based model for Japanese zero anaphora resolution that simultaneously identifies an appropriate case frame for a given predicate. We adopt a log-linear model, and utilize lexical features obtained from large-scale lexicalized case frames, which are automatically constructed from the Web, as well as non-lexical features that represent syntactic preferences, which are obtained from relatively small training data.

1. はじめに

照応とは，自然言語テキスト中のある表現 (照応詞) が他の表現 (先行詞) を指し示す現象のことで，照応解析は計算機によるテキスト理解を目指す上で重要な技術である．日本語においては照応詞が省略されるゼロ照応と呼ばれる現象が頻出し，高精度なゼロ照応解析システムの実現は情報抽出や質問応答，機械翻訳などといった多くの自然言語処理アプリケーションの高度化に必要不可欠であると言える．

(i) 彼は悪戯が好きで先生も (ϕ ニ) 手を焼いている．

例えば上記のような文では,「焼く」のニ格が省略されており，省略されたニ格 (ゼロ代名詞) は文頭の「彼」を照応している．一般に「焼く」のニ格が《人》となることは多くないが，この場合，ヲ格が「手」であることからニ格が「彼」であると推測できる．このようにゼロ照応解析は，対象用言の他の格要素と密接に関連しており，このため述語項構造解析の一部として解かれることが多い[1),2)]．

また，ゼロ照応解析の手掛かりは大きく語彙的知識と構文的知識に分けられる．このうち語彙的知識とは (i) において「焼く」のヲ格が「手」の場合，ニ格は《人》となることが多いなどといった知識のことで，各用言の各意味ごとに必要となるため人手で記述することは困難であり，大量のテキスト等から自動獲得する必要がある．一方，構文的知識とは，(i) における「彼」のように副助詞「は」を伴なって出現した語や文頭に出現した語は先行詞となりやすいなどといった知識のことで比較的小規模の関係タグ付きコーパスから獲得することが可能である．

我々はこのような考えから，自動獲得した大規模格フレームに基づくゼロ照応解析の確率モデル[3)] を提案した．このモデルでは，まず，共参照関係の認識を行いテキスト中に出現する談話要素を抽出した後，解析対象の用言に対し考えられる格フレーム，および，それぞれの格フレームと談話要素の考えられるすべて対応付けを確率的に評価し，もっとも確率が高いと判断された格フレーム，対応付けをその用言の解析結果としている．確率的評価を行う際は，独立性を仮定し，複数の語彙的選好を表す部分と構文的選好を表す部分に分解し，それぞれの確率の積を取ることで最終的な確率的評価値を求めている．しかしながら，このモデルは，独立性を仮定して種々の選好を表す部分に分解しているため，新たな素性，特に既存の素性と重複するような素性を導入が困難であるという問題点があった．

本稿では，ゼロ照応解析を対数線形モデルを使ってモデル化し，従来のモデルでは自然に導入できなかった意味クラス[4)] に関する素性や顕現性に関する素性を導入したモデルを提案する．また，素性ごとの重みや，各素性を使用しなかった場合の解析精度の変化を調べることにより，格ごとに有効となる素性の傾向を明らかにする．

†1 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University

Vol.2010-NL-198 No.9
2010/9/17

表 1 「焼く」の格フレーム

格スロット		用例 (頻度)	汎化用例 (割合)
焼く (1)	ガ格	私:39, 主人:26, 娘:22,	CT-人:0.620, NE-人:0.116,
		母さん:19, ···, 麻子:2, ···	CL-887:0.070, ···
	ヲ格	パン:9265, ケーキ:4495,	CT-食べ物:0.711, CL-883:0.221,
		肉:4057, 魚:2002, ···	CT-動物部位:0.105, CT:動物:0.077, ···
	ニ格	おやつ:35, 朝食:33,	CT-抽象物:0.233, CT-食べ物:0.171,
		誕生日:29, 土産:21, ···	CL-624:0.076, ···
	デ格	フライパン:894, 中火:425, ···	CT-人工物:0.356, CL-291:0.252, ···
			⋮
焼く (2)	ガ格	誰:7, 先生:7, 皆:5,	CT-人:0.372, NE-人:0.128,
		家族:4, 政府:3, 全員:3, ···	CT-組織・団体:0.128, ···
	ヲ格	手:6864	
	ニ格	加減:70, 子供:52, 攻撃:43,	CT-抽象物:0.432, CT-人:0.172,
		扱い:40, 対策:32, 娘:30, ···	NE-人:0.060, ···
			⋮
焼く (3)	ガ格	俺:1, 夫:1	
	ヲ格	ファイル:20, 曲:14, 音楽:9, ···	CT-抽象物:0.645, CT-人工物:0.273
	ニ格	CD:3106, DVD:2066, ···	CL-70:0.829, ···
	デ格	ライティングソフト:10, ···	CL-抽象物:0.294, CT-人工物:0.191, ···
			⋮

2. 大規模格フレームと用例の汎化

格フレームとは，用言とそれに関係する名詞を用言の用法ごと，格ごとに整理した知識であり，テキストから自動構築する手法が提案されている[5]．表 1 に「焼く」の格フレームの例を示す.「《食べ物》を焼く」,「《手》を焼く」,「《データ》を焼く」という意味ごとに格フレームが構築されており，それぞれの用法において取り得る格の種類と，その格の用例が記されている．

近年，Web から大量のテキストを容易に収集できるようになったことから大規模な格フレームの構築が可能となってきているが，Web テキスト 16 億文から構築した格フレームであってもゼロ照応解析を行う上で十分なカバレッジを持っているとは言えない[6]．例えば，「甥」や「太郎」などといった語は表 1 に示す格フレーム『焼く (1)』のガ格を埋めることができると考えられるが，これらの用例は『焼く (1)』のガ格には含まれていない．従って，より高精度なゼロ照応解析の実現のためにはこれらの語と『焼く (1)』のガ格を対応付けられるようにする必要があると考えられる．

このため，我々は形態素解析システム JUMAN6.0⋆1 の辞書に付与されているカテゴリ情報 (CT)，および，IREX[7] で定義された固有表現情報 (NE) を用いて格フレームの用例の汎化を行い，ゼロ照応解析に利用してきた[3]．ここで，JUMAN のカテゴリ情報とは普通名詞に付与される情報で,「私」や「主人」などに付与されている ‘人’ や,「パン」や「ケーキ」などに付与されている ‘人工物-食べ物’ など計 22 種類のカテゴリが定義されており，各カテゴリが付与された名詞が用例中にどのくらい含まれているかという情報を格フレームに付与している．また，IREX で定義された固有表現情報は，‘PERSON(人)’, ‘ORGANIZATION(組織・団体)’, ‘LOCATION(場所)’, ‘ARTIFACT(人工物)’, ‘DATE(日付)’, ‘TIME(時間)’, ‘MONEY(金額)’, および，‘PERCENT(割合)’ の 8 種類で，事前に固有表現認識を行ったテキストから格フレームを構築することにより，各格スロットの用例に占める対象の固有表現であると解析された用例の割合を格フレームに付与している．

本研究では，さらに風間らが導入した確率的クラスタリング法[4] によりクラスタリングされた意味クラス (CL) を用いた格フレームの用例の汎化を行う．本研究では 100 万個の名詞を対象に 2,000 クラスに分類した結果得られた意味クラスを使用する．以下に意味クラスと，それぞれのクラスへの帰属度上位の名詞の例を示す．括弧内の数字はその名詞の意味クラスへの帰属度を表している．

CL-70 CD(0.896), DVD(0.837), CDROM(0.603), カセットテープ (0.512), ···

CL-291 中弱火 (0.720), とろ火 (0.715), 中火 (0.681), 遠火 (0.678), ···

CL-624 夕飯 (0.926), 夕食 (0.925), 昼食 (0.882), 朝食 (0.868), ···

CL-884 ラーメン (0.860), うどん (0.801), カレー (0.793), ケーキ (0.749), ···

CL-887 母 (0.909), 両親 (0.875), 母親 (0.838), 夫 (0.775), 父親 (0.774), ···

格フレームに汎化情報として記述する際は，各クラスへの帰属度が 0.001 以上である名詞を対象とし，用例中の割合と帰属度の重み付け平均を計算し，意味クラスに関する情報として記述する．例えば，クラス ‘CL-70’ への帰属度が 0.896 である CD を 3,106 個，0.837 である DVD を 2,066 個，帰属度が 0.001 以下である名詞を 271 個含む格スロットがあった場合，次式により ‘CL-70:0.829’ という情報を意味クラスに関する情報として付与する．

$$(0.896 \times 3106 + 0.837 \times 2066)/(3106 + 2066 + 271) \approx 0.829 \quad (1)$$

表 1 に示す格フレームの汎化用例において ‘CT’, ‘NE’, ‘CL’ で始まる情報がそれぞれカテゴリ，固有表現，意味クラスによる汎化情報を表している．

⋆1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

Vol.2010-NL-198 No.9
2010/9/17

3. 識別モデルに基づくゼロ照応解析

3.1 提案モデルの概要

ゼロ照応解析は基本的に先行研究[3)]と同様の手順で行う．以下に解析手順の概要を示す．

(1) 形態素解析，固有表現認識，構文解析を行う[⋆1]．

(2) 共参照解析を行い，テキスト中に出現した談話要素を認識する．

(3) 入力テキスト 1 文ごとに，文末の用言から順に以下の手順で述語項構造を決定する．

- (a) 解析対象用言の格フレームを 1 つ選ぶ．
- (b) 解析対象用言と直接係り受け関係にある語と格スロットの対応付けを行う．
- (c) 対応付けられなかったガ格，ヲ格，ニ格の格スロットと，対象用言の格スロットと対応付けられていない談話要素の対応付けを行う．
- (d) すべての格フレーム，および，格フレームと談話要素の対応付けを確率的に評価し，もっとも確率が高いと判断された対応付けを解析結果とする．

先行研究[3)]と異なる点は，手順 (3d) における確率的評価方法のみである．本研究では次節で述べるような，識別モデルにより対応付けの評価を行う．

3.2 識別モデルに基づく対応付けの評価

入力テキスト t，解析対象用言 p が与えられたとき，格フレーム cf，格フレームと談話要素の対応付け a の条件付き確率を，以下のような対数線形モデルで表現する．

$$P(cf, a|p, t; \Lambda) = \frac{1}{Z(p,t)} \exp\{\Lambda \cdot \boldsymbol{F}(cf, a, p, t)\} \tag{2}$$

$$Z(p,t) = \sum_{\{cf,a\} \in C(p,t)} \exp\{\Lambda \cdot \boldsymbol{F}(cf, a, p, t)\} \tag{3}$$

ここで，$\boldsymbol{F}$ は次節で定義する素性関数で構成されるベクトル，Λ は素性関数に関する重み，$C(p,t)$ は与えられた用言が取り得る格フレーム，および，格フレームと談話要素の対応付けの集合を返す関数である．

対応付けの評価は各格ごとに行うのではなく，格フレーム選択や非省略要素の格フレームとの対応付けも含めた述語項構造解析全体で行う．例えば，1 節の例文 (i) 中の「焼く」の解析を考える場合，以下に挙げるような様々な対応付けを考え，その中で最も確率的評価の高い対応付けを最終的な述語項構造として決定する．

- 〈格フレーム:『焼く (1)』, ガ格:先生, ヲ格:手, ニ格:彼〉
- 〈格フレーム:『焼く (1)』, ガ格:先生, ヲ格:手, ニ格:(対応なし) 〉
- 〈格フレーム:『焼く (1)』, ガ格:(対応なし), ヲ格:手, ニ格:先生〉
- 〈格フレーム:『焼く (1)』, ガ格:彼, ヲ格:手, ニ格:(対応なし) 〉
- 〈格フレーム:『焼く (2)』, ガ格:先生, ヲ格:手, ニ格:彼〉
- 〈格フレーム:『焼く (2)』, ガ格:先生, ヲ格:手, ニ格:(対応なし) 〉

...

3.3 パラメータ推定

パラメータ推定は，N インスタンスの学習セット $\{(a^{(1)}, p^{(1)}, t^{(1)}), (a^{(2)}, p^{(2)}, t^{(2)}), \ldots, (a^{(N)}, p^{(N)}, t^{(N)})\}$ が与えられた時，事後確率を最大化するような格フレーム cf とパラメータ Λ の組み合わせを推定することにより行う．

$$\max_{CF,\Lambda} \left\{ \sum_{n=1}^{N} \log P(cf^{(n)}, a^{(n)}|p^{(n)}, t^{(n)}) - C||\Lambda||^2 \right\} \tag{4}$$

ここで，$CF = \{cf^{(1)}, cf^{(2)}, \ldots cf^{(N)}\}$ は各インスタンスの取り得る格フレームの集合のベクトルを表す．一般にゼロ照応関係が付与されたデータは格と先行詞の対応付けのみが付与されており，適切な格フレームに関する情報は付与されていない．このため，通常の最大エントロピー法と異なり，各インスタンスに対し事後確率が最大とするような格フレーム $cf^{(n)}$ を求める必要がある．パラメータ推定の手順を以下に示す．

(1) パラメータ Λ を初期化する．

(2) 各インスタンスごとに，現在のパラメータ Λ で $P(cf, a|p, t; \Lambda)$ が最大となるよう，すわなち，以下の式を満たすように $cf^{(n)}$ を更新する．

$$cf^{(n)} = \underset{cf}{\operatorname{argmax}}\, P(cf^{(n)}, a^{(n)}|p^{(n)}, t^{(n)}; \Lambda) \tag{5}$$

すべてのインスタンスに対し，$cf^{(n)}$ が更新されなかった場合，現在のパラメータ Λ を最終的な推定値とする．

(3) $cf^{(n)}$ が更新された場合は，N インスタンスの学習セット $\{(cf^{(1)}, a^{(1)}, p^{(1)}, t^{(1)}), (cf^{(2)}, a^{(2)}, p^{(2)}, t^{(2)}), \ldots, (cf^{(N)}, a^{(N)}, p^{(N)}, t^{(N)})\}$ に対し，確率モデルの事後確率

⋆1 本研究では，形態素解析には JUMAN6.0，構文解析には KNP3.0 を使用する．

を最大化するように Λ を更新し，(2) に戻る．本研究では過学習を防ぐため L_2 正則化を行い，以下の式を最大化するパラメータ Λ を求める[⋆1]．

$$\mathcal{L}_\Lambda = \sum_{n=1}^{N} \log P(cf^{(n)}, a^{(n)}|p^{(n)}, t^{(n)}) - C||\Lambda||^2 \quad (6)$$

式 (6) を最大化する Λ は L-BFGS 法[8] により求める[⋆2]．また，上記の手順は手順 (2)，手順 (3)，いずれにおいても尤度は単調に増加することから収束は保証されるが，式 (4) の中括弧の中身を最大化することは保証しない．しかしながら，複数の初期値を用いて上記の手順を行った結果，いくつかの初期値で同じ最大値に収束したことから，複数の初期値を用いてその中で最大の値に収束した場合のパラメータを用いることで，多くの場合，最適なパラメータ Λ を求められると考えられる．

パラメータに推定に用いるインスタンスは訓練データに出現した全用言から生成する．例えば 1 節の例文 (i) が与えられた場合，{(a:［ガ:先生, ヲ:手, ニ:彼］, p:焼く, t:「彼は ··· 焼いている.」), (a:［ガ:悪戯］, p:好き, t:「彼は ··· 焼いている.」)} という 2 つのインスタンスが生成される．この例からも分かるようにゼロ照応を含まない述語項構造からもインスタンスを作成する．

3.4 使用する素性

各用言ごとに 1 つ設定する素性 1 個と，格解析対象の 3 つの格，すわなち，ガ格，ヲ格，ニ格ごとに個別に設定する 8 種の素性の計 9 種 25 個の素性を使用する．使用する素性は，顕現性に関する素性と，ゼロ代名詞として先行詞と対応付けられたかに関する素性以外は実数素性である．

格解析スコア　河原ら[9] が提案した格フレームに基づく構文・格解析の統合的モデルにより与えられるゼロ照応を除いた述語項構造解析のスコア．省略格を除く対応付けがどのくらい確からしいかを表す．各用言に対し 1 つ設定する．

以下で導入する素性は各格ごとに設定する．また，いずれの素性もあくまで対象の格が直接係り受け関係にある語によって埋められなかった場合にのみ考慮する素性である．例えば，3.2 節の最後に挙げた対応付け〈格フレーム:『焼く (1)』, ガ格: 先生, ヲ格:手, ニ格:彼〉であれば，ガ格，ヲ格は直接係り受け関係にある語により埋められており，これらの対応付けに関する情報は格解析スコアに含まれていると考えられることから，ガ格，ヲ格に関する素性はいずれも 0 とし，ニ格に関する素性のみ考慮する．また,〈格フレーム:『焼く (1)』, ガ格:先生, ヲ格:手, ニ格:(対応なし) 〉という対応付けの場合，ニ格に関しても先行詞と対応付けられていないため，対象の格の埋まり易さに関する素性を除いてニ格に関する素性も 0 となる．

用例 PMI　対象の格スロットの用例と，対応付けられた先行詞の自己相互情報量 (PMI)．対象の格スロットにおける先行詞の占める割合を，一般のテキスト中における先行詞の占める割合で割ることにより計算する．先行詞となる談話要素が複数の語によって言及されている場合は，それらの中で最大となるものを使用する．

対象の格スロットが先行詞に対応付けられていない場合は 0，また，対象の格スロットの用例に先行詞が含まれていない場合は一定の小さな値 γ[⋆3]を与える．

意味クラス PMI　対象の格スロットの意味クラス情報と，対応付けられた先行詞の意味クラスの自己相互情報量．先行詞が帰属度 0.001 以上となる意味クラスを持っている場合のみ使用する．

カテゴリ PMI　対象の格スロットのカテゴリ情報と，対応付けられた先行詞のカテゴリの自己相互情報量．先行詞にカテゴリが付与されている場合のみ使用する．

固有表現 PMI　対象の格スロットの固有表現情報と，対応付けられた先行詞の自己相互情報量．先行詞が固有表現であると解析されている場合のみ使用する．

用例 PMI，意味クラス PMI，カテゴリ PMI，固有表現 PMI の 4 種の素性はいずれも先行詞がどのくらい対象の格スロットに対応付けられやすいかを表す素性である．これらの素性は基本的に大量のタグなしコーパスから獲得した知識から計算する．

先行詞の出現位置と対象格の PMI(位置 PMI)　先行詞の出現位置と省略対象格の自己相互情報量 (PMI)．“係り元用言のヲ格”，“係り元用言の省略されたヲ格”，“1 文前の文末”，“2 文前のヲ格” などといったように，どのような格を伴なって出現したか，ゼロ代名詞として出現したかどうかなどの情報も含む 80 程度の位置カテゴリを設定し，事前に訓練データから計算する．

例えば訓練データ中のヲ格ゼロ代名詞の総数が 127，先行詞候補の総数が 21,685，ある位置カテゴリ l に存在する先行詞候補の数が 58，そのうち先行詞となるものが 12 あ

⋆1 本稿における実験では $C = 10$ とした．
⋆2 libLBFGS 1.9 (http://www.chokkan.org/software/liblbfgs/) を使用した．
⋆3 本稿の実験では $\gamma = \log(0.000001)$ とした．

る場合，以下の式により位置カテゴリ l とヲ格の PMI は 3.56 と計算される．

$$PMI(\text{ヲ格}, l) = \log\left(\frac{P(\text{ヲ格}, l)}{P(\text{ヲ格})P(l)}\right) = \log\left(\frac{\frac{12}{21,685}}{\frac{127}{21,685}\cdot\frac{58}{21,685}}\right) = 3.56 \qquad (7)$$

表 2 にヲ格ゼロ代名詞の先行詞となりやすい位置カテゴリの例と PMI を示す．ひとつの先行詞が複数の位置カテゴリに該当した場合は PMI が最大となるものを使用する．

先行詞の顕現性　談話要素の顕現性を照応関係を解く上で有力な手掛りとなる[10),11)] ことから顕現性に関する素性を導入する．

具体的には，以下のルールで談話要素の顕現性を計算し，先行詞候補の顕現性が 1 以上である場合，この素性を 1，そうでない場合 0 とする．

- +2.0: 文末，または，副助詞「は」を伴なって出現
- +1.0: 読点，または，格助詞「が」「を」を伴なって出現
- +1.0: ゼロ代名詞の照応先となる
- ×0.5: 文区切りを通過

先行詞の出現位置と対象格の PMI と先行詞の顕現性の 2 種の素性は，先行する解析結果によって変化する素性であり，実験において正しいゼロ照応関係を使用する場合と，先行する自動解析結果を用いる場合で精度が異なる要因となる素性である．

対象の格の埋まり易さ　対象の格スロットが直接係り受けにある語によって埋められる確率，すなわち，河原ら[9)] の格フレームに基づく構文・格解析の統合的モデルにおける格スロット生成確率に関する素性．対象の格スロットが直接係り受けにある語によって埋められる確率に関する素性ではあるが，あくまで対象の格スロットが直接係り受けにある語によって埋められない場合にのみ考慮する．

対象の格がゼロ代名詞として先行詞と対応付けられた場合は対象の格スロットが直接係り受けにある語によって埋められる確率を，対応付けられなかった場合は，その値を 1 から引いた値，すわなち，対象の格スロットが直接係り受けにある語によって埋められない確率を素性として使用する．

対象の格が対応付けられたかどうか　対象の格がゼロ代名詞として先行詞と対応付けられた場合に 1，それ以外の場合に 0 とする．各格がどのくらい埋まり易いかを調整するパラメータとなる．

表 2　ヲ格ゼロ代名詞の先行詞になりやすい位置カテゴリ

位置カテゴリ	PMI
後続する並列用言の省略されたヲ格	4.22
係り元用言のヲ格	3.56
先行する並列用言のヲ格	3.38
係り先用言の係り先用言のヲ格	3.00
係り先用言の省略されたヲ格	2.94
...	
係先用言のニ格	2.43
...	
係り先用言の省略されたガ格	2.22
...	
前文の文頭	1.59
...	

4. 実　　験

4.1 使用するデータ

照応タグ付きデータとして，Web テキスト 186 記事に京都テキスト[12)] と同様の基準で照応関係タグを付与したデータ (Web コーパス)，および，NAIST テキストコーパス[13)] を使用する．

Web コーパスを使用する際は，124 記事を訓練用に 62 記事をテストに使用する．NAIST テキストコーパスを使用する際は Iida ら[14)] と同様の 137 記事を訓練に 150 記事をテストに使用する．ただし，本研究では受け身，および，使役の場合であっても表層格の解析を行っているのに対し，NAIST テキストコーパスでは原形に戻した場合の格が付与されているため，受け身，および，使役形である用言は除いて使用する．

述語項構造解析以外の解析結果が原因となる解析誤りを除くため，形態素情報，固有表現情報，係り受け情報，共参照関係はコーパスに付与された正しい情報・関係を使用する．また，3.4 節で導入した先行詞の出現位置と対象格の PMI，および，先行詞の顕現性の 2 種の素性は，先行する解析結果によって変化する素性となっていることから，先行するゼロ照応解析に関しては，コーパスに付与された正しい関係を使用する条件と，自動解析の結果を使用する条件の 2 条件で実験を行う．

4.2 実験結果と考察

Web コーパスを用いた実験の結果を表 3 に示す．各素性の有効性を確かめるため素性を 1 種ずつ除いた実験の結果も示している．また，すべての素性を用いた場合の学習された各

表 3 Web コーパスを用いた実験の結果

使用しない素性	正しいゼロ照応解析関係を使用 再現率	適合率	F 値	先行する自動解析結果を使用 再現率	適合率	F 値
すべての素性を利用	0.325 (79/243)	0.585 (79/135)	0.418	0.255 (62/243)	0.521 (62/119)	0.343
格解析スコア	0.206 (50/243)	0.303 (50/165)	0.245	0.144 (35/243)	0.660 (35/53)	0.236
用例 PMI	0.321 (78/243)	0.624 (78/125)	0.424	0.210 (51/243)	0.537 (51/95)	0.302
意味クラス PMI	0.325 (79/243)	0.590 (79/134)	0.419	0.230 (56/243)	0.483 (56/116)	0.312
カテゴリ PMI	0.296 (72/243)	0.581 (72/124)	0.392	0.198 (48/243)	0.403 (48/119)	0.265
固有表現 PMI	0.321 (78/243)	0.582 (78/134)	0.414	0.210 (51/243)	0.432 (51/118)	0.283
位置 PMI	0.251 (61/243)	0.610 (61/100)	0.356	0.165 (40/243)	0.500 (40/80)	0.248
顕現性	0.308 (75/243)	0.573 (75/131)	0.401	0.259 (63/243)	0.534 (63/118)	0.349
対象格の埋まり易さ	0.317 (77/243)	0.626 (77/123)	0.421	0.218 (53/243)	0.491 (53/108)	0.302
位置 PMI, 顕現性	0.156 (38/243)	0.521 (38/73)	0.241	0.156 (38/243)	0.521 (38/72)	0.241

表 4 学習された各素性に対する重み

素性	重み		
格解析スコア		**1.300**	
	ガ格	ヲ格	ニ格
用例 PMI	0.185	0.199	**0.214**
意味クラス PMI	0.041	**0.206**	-0.006
カテゴリ PMI	**0.733**	0.211	0.195
固有表現 PMI	**0.261**	0.132	-0.127
位置 PMI	**0.869**	0.622	0.400
顕現性	**1.242**	0.433	0.596
対象格の埋まり易さ	0.024	0.815	**0.994**
対象格が先行詞と対応付けられたか	**-3.033**	-2.053	-2.570

太字は各行でもっとも絶対値が大きいことを示す.

素性に対する重みを表 4 に示す.

すべての素性を用いた場合の精度を見ると，再現率に比べて適合率が高くなっていることが分かる．これは提案モデルにおいてパラメータを学習する際に生成するインスタンスの多くはゼロ照応を含んでいないため，ゼロ照応を含まないような対応付けを出力しやすいパラメータが学習されたためであると考えられる．実際，訓練に用いた 126 記事から生成されるインスタンス 1,367 個中，ゼロ照応を含むインスタンスは 384 個のみであった．

この傾向は表 4 において，対象格が先行詞と対応付けられたかどうかに関する素性に大きな負の重みが与えられていることからも確認できる．対象格が先行詞と対応付けられた場合に 1 となる素性に大きな負の重みが与えられていることから，このような対応付けが選択されるためには，カテゴリ PMI や位置 PMI などの他の素性の値が十分に大きな値となることが必要となる．

表 3 の最後に示したのは，先行する解析結果によって変化する素性である位置 PMI と顕現性素性をともに使用しなかった場合の精度である．これらの素性を使わなかった場合，先行する解析結果による影響を受けないため，正しいゼロ照応解析関係を使用した場合と自動解析結果を使用した場合の精度が一致することが確認できる．

次に，学習された各素性に対する重みに注目する．まず，ガ格は用例 PMI や意味クラス PMI に比べ，カテゴリ PMI や固有表現 PMI に大きな重みが割り当てられている．これはある表現がガ格を埋められるかどうかは，特定の形態素であるかどうかよりも,《人》であるか《組織》であるかなどといった大きなクラスに関する情報が重要であることを示していると考えられる．一方，ヲ格に対しては意味クラス PMI に比較的大きな重みが割り当てられており，このことからあるヲ格を埋めることができる表現の分布はガ格の場合より限定的であることが多いことが推測される．

また，位置 PMI や顕現性に対する素性の重みはガ格が一番大きいのに対し，対象格の埋まりやすさに関する素性の重みはガ格が際だって小さくなっている．これは，ヲ格やニ格の場合，その格が埋まりやすい用言と埋まりにくい用言が存在し，そのような語彙的情報がゼロ照応においても重要な手掛かりとなるのに対し，ガ格が埋まるかどうかはどのような用言であるかといった語彙的情報よりも先行詞の出現位置や省略要素の顕現性などの影響が強いという傾向を表していると考えられる．

最後に NAIST テキストコーパスを用いた実験の結果を表 5 に示す．Web コーパスを用いた実験とほぼ同程度の精度が得られたが，正しいゼロ照応解析関係を使用した場合と先行する自動解析結果を使用した場合の精度の差はやや大きくなっている．NAIST テキストコーパスを用いてゼロ照応解析の評価を行っている他の研究と比較すると，本研究では受け身，使役形となっている用言を解析対象から除いているなど条件がやや異なるため単純には比較できないものの，Iida ら[14] よりは低い精度，Imamura ら[2] と同程度の精度であると考えられる．

表 5 NAIST テキストコーパスを用いた実験の結果

正しいゼロ照応解析関係を使用			先行する自動解析結果を使用		
再現率	適合率	F 値	再現率	適合率	F 値
0.405 (408/1008)	0.466 (408/876)	0.433	0.288 (290/1008)	0.351 (290/827)	0.316

5. おわりに

本稿では，ゼロ照応解析を対数線形モデルを使ってモデル化し，意味クラスに関する素性や顕現性に関する素性を導入したモデルを提案した．また，素性ごとの重みや，各素性を使用しなかった場合の解析精度の変化を調べることにより，格ごとに有効となる素性の傾向を明らかにした．今後，用言対の項共有情報や，用言間の解析結果の整合性など，現在考慮していない素性をモデルに取り込みさらなる精度の向上を目指す予定である．

参 考 文 献

1) Taira, H., Fujita, S. and Nagata, M.: A Japanese Predicate Argument Structure Analysis using Decision Lists, *Proc. of EMNLP'08*, pp.523–532 (2008).
2) Imamura, K., Saito, K. and Izumi, T.: Discriminative Approach to Predicate-Argument Structure Analysis with Zero-Anaphora Resolution, *Proc. of ACL-IJCNLP'09*, pp.85–88 (2009).
3) Sasano, R., Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Zero Anaphora Resolution, *Proc. of COLING'08*, pp.769–776 (2008).
4) Kazama, J. and Torisawa, K.: Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations, *Proceedings of ACL-08: HLT*, pp.407–415 (2008).
5) 河原大輔，黒橋禎夫：格フレーム辞書の漸次的自動構築，自然言語処理， Vol.12, No.2, pp.109–131 (2005).
6) Sasano, R., Kawahara, D. and Kurohashi, S.: The Effect of Corpus Size on Case Frame Acquisition for Discourse Analysis, *Proc. of NAACL-HLT'09*, pp.521–529 (2009).
7) IREX 実行委員会（編）：IREX ワークショップ予稿集 (1999).
8) Nocedal, J.: Updating Quasi-Newton Matrices with Limited Storage, *Mathematics of Computation*, Vol.35, No.151, pp.773–782 (1980).
9) 河原大輔，黒橋禎夫：自動構築した大規模格フレームに基づく構文・格解析の統合的モデル，自然言語処理， Vol.14, No.3, pp.67–81 (2007).
10) Lappin, S. and Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, Vol.20, No.4, pp.535–562 (1994).
11) Mitkov, R., Evans, R. and Orăsan, C.: A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method, *Proc. of CICLing'02* (2002).
12) 河原大輔，黒橋禎夫，橋田浩一：「関係」タグ付きコーパスの作成，言語処理学会第 8 回年次大会発表論文集，pp.495–498 (2002).
13) Iida, R., Komachi, M., Inui, K. and Matsumoto, Y.: Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations, *Proc. of ACL'07 Workshop: Linguistic Annotation Workshop*, pp.132–139 (2007).
14) 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, p.Article 12 (2007).