ゼロ代名詞照応付き述語項構造解析の対話への適応

今村 賢治　　東中 竜一郎　　泉 朋子
日本電信電話株式会社，NTT メディアインテリジェンス研究所
{imamura.kenji,higashinaka.ryuichiro,izumi.tomoko}@lab.ntt.co.jp

1　はじめに

意味役割付与 (semantic role labeling; SRL) または述語項構造解析 (predicate-argument structure analysis) は，文から「誰が何をどうした」情報を得るための重要な解析技術の一つである。従来これらは，コーパスが新聞記事であるなどの理由で，書き言葉で多く研究されてきた (Màrquez et al., 2008; 松林他, 2013)。

一方，近年のスマートフォンの普及に伴い，Apple 社の Siri，NTT ドコモ社のしゃべってコンシェルなど，音声による人とコンピュータの対話システムが身近に使われ始めている。人・コンピュータの対話システムを構築するためには，人間の発話を理解し，システム発話とともに管理する必要があるが，対話理解に対しても，述語項構造は有効なデータ形式であると考えられる。しかし，新聞記事と対話では，発話人数，口語の利用，文脈など，さまざまな違いがあるため，既存の新聞記事をベースとした述語項構造解析を対話の解析に利用した際の課題は，不明な点が多々ある。たとえば，日本語対話ではゼロ代名詞がごく自然に出現するので，述語項構造解析にはゼロ代名詞照応処理も必要となる。

A:　$[iPad\ Air]_{ga}$ がほしい。
B:　いつ $(\phi)_{ga}$ $(\phi)_{o}$ 買うの*?*

本稿では，新聞記事解析用に提案されたゼロ代名詞照応機能付き述語項構造解析を，日本語の雑談対話に適用する。適用の際には，新聞記事から対話への一種のドメイン適応とみなす。意味役割付与 (SRL) のドメイン適応 (Pradhan et al., 2008) では，適応に必要な要素として，未知語対策とパラメータ分布の違いの吸収を挙げている。本稿では，パラメータ分布の違いに焦点を当て，新聞記事用より高精度な対話用の述語項構造解析を構築する。

2　雑談対話の特徴

まず我々は，2 名の参加者による雑談対話を収集し，その対話に述語項構造データの付与を行った。雑談対話は，参加者にテーマ（話題）だけを示し，キーボード対話形式で収集した。したがって，音声対話に含まれるようなフィラーや繰り返しは少ない。参加者に提示した話題は，食事，旅行，趣味テレビ・ラジオなど，20 ジャンルのうちの一つである。雑談対話と，その述語項構造アノテーションの例を図 1 に示す。

表 1: コーパスサイズ

コーパス	セット	記事/対話数	文/発話数	述語数
NAIST	訓練	1,751	24,225	67,142
	開発	480	4,833	13,594
	テスト	695	9,272	25,497
雑談対話	訓練	184	6,960	7,470
	テスト	101	4,056	5,333

述語項構造アノテーションは，毎日新聞をベースにしている NAIST テキストコーパス (Iida et al., 2007) に準拠する形で行った。ただし，NAIST コーパスでは，先行詞が記事内に現れない「外界照応」は 1 種類しか定義されていないが，対話の場合，一人称・二人称代名詞が省略されることも多いため，外界照応を，exo1(一人称)，exo2(二人称)，exog(その他の外界照応）の 3 種類に細分した (松林他, 2013)。

今回作成した雑談対話コーパスと，NAIST テキストコーパスの概要を表 1 に示す[1][2]。対話コーパスは，NAIST コーパスの約 1/10 のサイズである。NAIST コーパスは，訓練，開発，テストに 3 分割したのに対し，対話コーパスは訓練とテストの 2 分割とした。

表 2 は，NAIST および対話コーパスの訓練セットにおける，項の分布を示したものである。各項は，その位置や文法関係により，以下の 7 分類した。

- **Dep:** 述語と項が直接の係り受け関係にある場合
- **文内ゼロ:** 述語と項が同じ文（発話）内にあるが，直接の係り受け関係がない場合
- **文間ゼロ:** 述語と項が異なる文にある場合
- **exo1/exo2/exog:** 項が記事（対話）内に存在しない場合。それぞれ，一人称ゼロ代名詞，二人称ゼ

[1]対話における「対話」と「発話」は，それぞれ新聞の「記事」「文」に相当するとみなす。
[2]NAIST コーパスの統計量は，1.4β を元にし，筆者らが文節化などの前処理を行った上で集計した。そのため，1.5 を用いた数値と一致していない。

A: 夏は (exo2)$_{ga}$ (exog)$_{ni}$ 出かけたりしましたか?
B: 8月は伊東の [花火大会*1]$_{ni}$ に (exo1)$_{ga}$ 行きました。
A: [花火*2]$_{o}$, [私*3]$_{ga}$ も見たかったです。
A: でも，今年は (exo1)$_{ga}$ 忙しくて (exo1)$_{ga}$ (*2)$_{o}$ 見に (exo1)$_{ga}$ (*2)$_{ni}$ 行けませんでした。

図 1: 雑談対話とその述語項構造アノテーションの例

表 2: 訓練セットにおける項の分布

格	コーパス	述語数	Dep	文内ゼロ	文間ゼロ	外界照応 exo1	exo2	exog	NULL
ga	NAIST	67,142	42.0%	29.7%	11.6%	0.0%	0.0%	16.4%	0.4%
	対話	7,470	28.0%	11.2%	12.6%	23.8%	5.6%	18.8%	0.0%
o	NAIST	67,142	33.8%	5.1%	1.3%	0.0%	0.0%	0.1%	59.8%
	対話	7,470	11.3%	3.8%	7.0%	0.2%	0.0%	3.1%	74.6%
ni	NAIST	67,142	16.2%	1.8%	0.4%	0.0%	0.0%	0.0%	81.6%
	対話	7,470	10.2%	2.5%	4.2%	0.7%	0.3%	10.1%	72.0%

ロ代名詞，それ以外を表す。exo1 および exo2 は，NAIST コーパスではアノテーションされていない。

- **NULL:** 項がこの述語では不必要な場合

まず，全述語の分布に着目すると，対話はすべての格で，直接係り受け (Dep) が減少している。それ以外の関係については，ガ格と，ヲ格ニ格で傾向が異なっている。

ガ格は，対話では文内ゼロ代名詞も減少し，減少分は一人称・二人称外界照応 (exo1, exo2) に割り当てられている。つまり，ガ格では，文内の項が減少し，ゼロ代名詞が新聞に比べて頻発する。ただし，その先行詞は一人称・二人称代名詞である可能性が高い。

ヲ格ニ格では，Dep の減少分は，文間ゼロ代名詞，またはその他の外界照応 (exog) に割り振られている。つまり，新聞記事では，ヲ格ニ格の大部分は述語と同じ文内に現れていたものが，対話では文外に現れることが多くなり，1 文に閉じない照応処理が重要となる。

3 基本方式と対話への適応

3.1 ゼロ代名詞照応付き述語項構造解析

本稿でベースとする述語項構造解析方法は，今村の方法 (Imamura et al., 2009) である。これは，NAIST コーパスを対象とした方法であるが，文内に存在する項，文間の項，外界照応を同時に決定できるという特徴がある。

処理は，記事（対話）全体を入力とし，各文（発話）ごとに以下のステップを実行する。

1. 入力文を形態素・構文解析する。なお，構文解析時には，文節とその主辞を特定しておく。なお，今回は，形態素情報は MeCab [3]で自動付与したが，構文情報は京都大学テキストコーパス 4.0[4] の情報を利用した。対話コーパスに関しては，CaboCha[5]で自動解析した構文木を使用した。

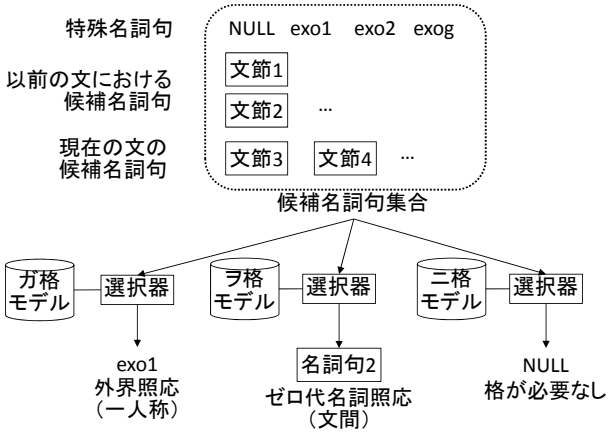

図 2: 提案方式の基本構成

2. 文から述語文節を特定する。これは，主辞が動詞，形容詞，形容動詞，名詞+助動詞「だ」の文節とし，品詞パターンで決定した。

3. 各述語について，述語の存在した文，およびそれより前方の文から，項の候補となる文節を取得する。具体的には，以下の文節が候補となる。

   - 文内の候補として，述語文節と係り受け関係にあるかどうかに関わらず，内容部が名詞句であるすべての文節を候補とする。
   - 文間の候補として，それまでに出現した文から，文脈的に項の候補となりうる文節を加える。詳細は 3.3 節で述べる。

[3] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[4] http://nlp.ist.i.kyoto-u.ac.jp/index.php?京都大学テキストコーパス

[5] http://code.google.com/p/cabocha/

表 3: 本稿で用いた平文コーパスから自動獲得した素性

種別	素性名	内容
述語	Frame	その格を必要とする場合 1，不必要な場合 0
両者の関係	$\log P(n\|c,v)$	述語 v，格 c から見た名詞句 n の生成確率 (実数)
	$\log P(v\|c,n)$	名詞句 n，格 c から見た述語 v の生成確率 (実数)
	$\log P(c\|n)$	名詞句 n から見た格 c の生成確率 (実数)

- 文章内に実体を持たない疑似候補として，外界照応 (exo1, exo2, exog) と，格を必要としない (NULL) を特殊名詞句として加える。

4. 述語文節，項の候補名詞句，両者の関係を素性化し，ガ，ヲ，ニ格独立に，候補名詞句からもっとも各格にふさわしい文節を選択する（図 2)。選択器のモデルは，最大エントロピー (ME) 法に基づくが，訓練時には候補名詞句集合全体で正規化することにより，ランキング学習を行っている。

3.2 素性

選択器で使用する素性に関しては，他の研究（たとえば (Gildea and Jurafsky, 2002)）と同様に，(1) 述語に関する素性，(2) 名詞句に関する素性，(3) 両者の関係に関する素性を使用する。なお，これらは名詞句の選択用モデルの素性であるので，名詞句の主辞に関する素性 Noun と，その他すべての二値素性を組み合わせたものも使用している。

また，対話用の素性として，述語に付随する機能表現 (Suffix 素性) と，述語の発話者と名詞句の発話者が同じかどうか (Speaker 素性) を含めた。未知語対策として，大規模平文コーパスから自動獲得した必須格情報 (Frame 素性) と係り受け言語モデル（3 種類）を，外部知識として使用し，素性の一部として選択器のモデルに組み込んだ（表 3)。

3.3 文脈処理

新聞記事のような書き言葉と対話では，明らかに文脈処理が異なると考えられ，本来なら，対話用の文脈管理を導入すべきである。しかし，対話システム全体から見た場合，文脈管理は述語項構造解析ではなく，システム･ユーザ発話を一括管理する対話管理モジュールに任せるべきであると考え，今回は新聞記事用と同じ文脈管理方法を使用する。なお，本稿の方式は，選択器に与える文間候補名詞句を取捨選択することによって，文間の文脈の制御を行っているので，候補名詞句を外部モジュールから陽に与えることで，文脈管理方法を変更することができる。

今回使用した文脈管理方法は，具体的には以下のとおりである。

- 対象述語の発話より以前の発話をさかのぼり，他の述語を含む発話（これを有効発話と呼ぶ）を見つける。これは，対話の場合，相槌など，述語を含まない発話が挿入されることがあり，これを無視するためである。

- 有効発話と対象述語の間に出現した全名詞句と，有効発話の述語で項として使われた名詞句（有効発話内の場合もあれば，それ以前の発話の名詞句の場合もある）を候補として加える。項として使われた名詞句は，その後も繰り返し使われることが多く，これに制限することで，効率的に候補制限することができるという観察結果に基づく (Imamura et al., 2009)。また，項として使われている限り，さかのぼる文数に制限がないため，広い文脈を見ることができる。

3.4 モデルパラメータの対話への適応

2 節で述べた，NAIST コーパスと対話コーパスの項分布の差異は，選択器のモデルパラメータをドメイン適応することで調整する。本稿では，モデルパラメータの適応手法として，素性空間拡張法 (Daume, 2007) を用いる。これは，素性空間を 3 倍に拡張することで，ソースドメインデータをターゲットドメインの事前分布とみなすのと同じ効果がある方法である。

具体的には，NAIST コーパスをソースドメインデータ，雑談対話コーパスをターゲットドメインデータとみなし，選択器の素性空間を拡張，モデルを学習する。選択器が項同定する際は，素性空間のうち，ターゲット空間と共通空間だけ用いる。この空間のパラメータは，ターゲットドメインに最適化されているだけでなく，ソースドメインだけに現れた素性も利用して項同定ができる。

4 実験

本節では，新聞ドメイン（NAIST コーパス)，対話ドメイン（対話コーパス）における述語項構造解析の精度を，パラメータ適応という観点から評価する。

評価に使用したコーパスは，表 1 に示したものである。また評価は，項ごとの適合率，再現率，F 値で評価した（外界照応も含む)。

比較した方式は，素性空間拡張によるドメイン適応を行った場合 (適応)，NAIST コーパスだけで訓練し

表 4: 雑談対話コーパステストセットにおける方式毎の精度

格	項のタイプ	項の数	適応 適合率	適応 再現率	適応 F 値	NAIST 訓練 適合率	NAIST 訓練 再現率	NAIST 訓練 F 値	対話訓練 適合率	対話訓練 再現率	対話訓練 F 値
ga	Dep.	1,575	86.0%	85.6%	**85.8**%	75.1%	88.8%	81.4%	85.8%	85.1%	85.5%
	文内ゼロ	747	62.0%	40.0%	48.7%	46.8%	51.3%	**48.9**%	58.5%	38.2%	46.2%
	文間ゼロ	767	36.2%	7.0%	11.8%	11.6%	9.0%	10.1%	38.0%	7.0%	**11.9**%
	exo1	1,193	61.2%	83.8%	**70.7**%	0.0%	0.0%	0.0%	59.8%	84.5%	70.0%
	exo2	281	69.7%	35.9%	**47.4**%	0.0%	0.0%	0.0%	61.4%	33.5%	43.3%
	exog	767	36.5%	64.5%	46.7%	21.6%	58.0%	31.5%	37.6%	63.9%	**47.4**%
	合計	5,330	61.8%	61.9%	**61.8**%	43.0%	43.1%	43.0%	61.4%	61.4%	61.4%
o	Dep.	585	84.7%	85.3%	**85.0**%	86.7%	73.3%	79.4%	82.0%	84.1%	83.0%
	文内ゼロ	178	51.5%	38.2%	**43.9**%	50.8%	16.9%	25.3%	48.9%	38.8%	43.3%
	文間ゼロ	399	40.5%	22.3%	28.8%	75.0%	0.8%	1.5%	43.4%	22.3%	**29.5**%
	exo1	19	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	exo2	7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	exog	98	21.4%	31.6%	25.5%	0.0%	0.0%	0.0%	32.0%	24.5%	**27.7**%
	合計	1,286	63.3%	53.4%	57.9%	82.8%	35.9%	50.1%	66.6%	52.4%	**58.4**%
ni	Dep.	554	87.6%	74.9%	**80.7**%	88.0%	65.2%	74.9%	88.3%	73.3%	80.1%
	文内ゼロ	82	43.5%	12.2%	19.0%	0.0%	0.0%	0.0%	50.0%	13.4%	**21.2**%
	文間ゼロ	169	34.1%	8.9%	14.1%	0.0%	0.0%	0.0%	33.3%	9.5%	**14.7**%
	exo1	32	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	exo2	4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	exog	265	36.9%	53.6%	**43.7**%	0.0%	0.0%	0.0%	39.1%	49.4%	**43.7**%
	合計	1,106	62.9%	52.6%	**57.3** %	88.0%	32.6%	47.6%	65.2%	51.0%	57.2%

た場合 (NAIST 訓練)，対話コーパスだけで訓練した場合 (対話訓練) である。雑談対話コーパステストセットでの結果を表 4 に示す。

まず，単独のコーパスで訓練した場合（NAIST/対話訓練）を比較すると，訓練セットとテストセットのコーパスが一致しないと精度が出ない。適応は，両者の良さをとり，合計では，対話テストセットのヲ格を除き，最高の F 値となった。

項のタイプごとの精度を見ると，特徴的なのは，表 4 のガ格の exo1/exo2 である。この 2 つは，ガ格の項のうちの約 28%を占めており，これが exo1 で 70.6%，exo2 で 47.3%の F 値で解析可能となった効果は大きい。

対話コーパスは，訓練セットのサイズが小さいにも関わらず，適応と対話訓練の精度がほぼ同じとなった。これは，対話コーパスのサイズが十分大きいという意味ではなく，適応が NAIST コーパスの知識を活かしきっていないものと思われる。対話コーパスを追加すれば，まだ精度が向上できる可能性がある。

いずれにしても，対話用述語項構造解析を構築するためには，少量でも対話の述語項構造アノテーションデータが効果があり，ドメイン適応は，新聞記事を対話に適応させるときにも有効である。

5 おわりに

本稿では，従来新聞記事で研究されていた述語項構造解析を，対話に適用した。対話と新聞記事では項の分布が異なるため，ドメイン適応技術を用いて，モデルパラメータを適応させた。結果，ドメイン適応を施すことにより，少ない対話コーパスからでも対話に頻出するゼロ代名詞を含む述語項構造解析ができるようになった。

今回は，パラメータ分布の差異に着目したが，ドメイン適応としては，語彙のカバレッジにも着目する必要がある。また，新聞と対話では明らかに文脈管理が異なる。文脈として，対話システムの発話管理を使ったときの有効性評価は今後の課題である。

参考文献

Hal Daume, III. 2007. Frustratingly easy domain adaptation. In *Proc. of ACL-2007*, pages 256–263.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Ryu Iida, Mamoru Komachi, Kentaro Inui, and Yuji Matsumoto. 2007. Annotating a Japanese text corpus with predicate-argument and coreference relations. In *Proc. of the Linguistic Annotation Workshop*, pages 132–139.

Kenji Imamura, Kuniko Saito, and Tomoko Izumi. 2009. Discriminative approach to predicate-argument structure analysis with zero-anaphora resolution. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pages 85–88.

Lluís Màrquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Computational Linguistics*, 34(2):289–310.

松林 優一郎, 飯田 龍, 笹野 遼平, 横野 光, 松吉 俊, 藤田 篤, 宮尾 祐介, 乾 健太郎. 2013. 日本語述語項構造アノテーションに関わる諸問題の分析. **情報処理学会研究報告** *2013-NL-214(12)*, pages 1–18, 11 月.