決定リストを用いた述語項構造解析

平　博順　藤田早苗　永田 昌明
NTT コミュニケーション科学基礎研究所
{taira, sanae}@cslab.kecl.ntt.co.jp　nagata.masaaki@lab.ntt.co.jp

1. はじめに

近年、意味解析の一つとして述語項構造解析が注目されている。述語項構造解析は、述語と項（日本語では述語と格関係にある名詞句）との関係を同定する技術である。表 1 に述語項構造のコーパスの一つである NAIST テキストコーパスの定義に基づく述語および事態性名詞に関する項構造の例を示す。NAIST テキストコーパスでは、テキスト中の述語および事態性名詞の基本形に対して意味的にガ格（主格）、ヲ格（対格、対象格）、ニ格（与格、目的格）となる名詞句を正解の項としてタグ付けがされている。「花子が太郎から与えられた本」というテキストがあった場合、述語「与えられた」の基本形「与える」に対してガ格、ヲ格、ニ格となる項はそれぞれ「太郎」、「本」、「花子」となる。また、NAIST テキストコーパスでは、通常の述語だけでなく事態性名詞にも項構造が付与されている。「花子からの電話で私は起きた」というテキストがあった場合、表面的には「電話」は単なる名詞であるが、意味的には「花子が（私に）電話する」という「事態 (event)」を含んでいると考えられる。このような事態を表す名詞を NAIST テキストコーパスでは事態性名詞と呼んでいる [5]。

このように、テキストから述語と名詞句の間の意味的な関係を自動抽出する技術は、さまざまな言語処理タスクにおいて有用であると考えられ、情報抽出 [1]、質問応答 [8] [10]、自動要約 [6] といったタスクにおいて、その利用が試みられてきている。

日本語の述語項構造コーパスとしては、GDA コーパス [15]、京都大学テキストコーパス Ver.4.0 [4]、NAIST テキストコーパス [3] などのコーパスが構築されているが、本稿ではこのうちタグ付けされた項の数が最大である NAIST テキストコーパスを対象に自動解析実験を行った。タスクとしては、英語の意味役割付与タスクにおけるコーパスである PropBank [9] と名詞化された動詞への意味役割付与コーパスである NomBank [7] に対する解析を合わせたものに似ている。しかし、意味役割付与タスクの場合は同一文内に閉じた解析である。NAIST テキストコーパスに対する解析では、文内だけでなく文外に存在する項も特定するゼロ代名詞解析 [2] [12] [16] も含まれたタスクになっているところが異なる。

表 **1.** 述語および事態性名詞に対する項構造

元テキスト：花子が太郎から与えられた本			
述語	項		
述語	ガ格 (主格)	ヲ格 (対格)	ニ格 (与格)
与える	太郎	本	花子
元テキスト：花子からの電話			
事態性名詞	項		
事態性名詞	ガ格 (主格)	ヲ格 (対格)	ニ格 (与格)
電話	花子	–	(私)

我々は、述語と事態性名詞の項構造解析をゼロ代名詞解析についても統一的に扱う方法で解析を行った。また、学習されたルールの可読性を重視して決定リストを用いたコーパスからの学習を行った。この決定リストは、様々な制約の下で、最も近くに位置する単語を探すルールが書かれたリストである。

本稿の構成は次の通りである。2 節で実験に用いた NAIST テキストコーパスについて説明し、3 節で我々の提案手法について述べる。4 節で実験結果について述べ最後に 5 節で結論を述べる。

2. NAIST テキストコーパス

NAIST テキストコーパスでは、毎日新聞 95 年版に関するタグ付けコーパスである京都大学テキストコーパス Ver3.0 をベースに、2929 記事、3 万 8384 文に対して、述語及び事態性名詞の基本形に対するガ格（主格）、ヲ格（対格）、ニ格（与格）の 3 つの必須格について正解項構造が付与されている。

表 2 に、我々がテストデータに用いた記事（4.1 節参照）における項と述語の係り受け関係別の出現頻度分布を示す。ここでは、係り受け関係を**係り受け関係有り** (項

表 **2.** 係り受け関係に対する項の分布

	述語		事態性名詞	
テストデータ全体	29,215	(100.0%)	6,836	(100.0%)
係り受け関係有り	21,051	(72.1%)	1,730	(25.3%)
同一文節内	225	(0.8%)	2,155	(31.5%)
文内ゼロ代名詞	5,086	(17.4%)	1,249	(18.3%)
文外ゼロ代名詞	2,853	(9.8%)	1,702	(24.9%)

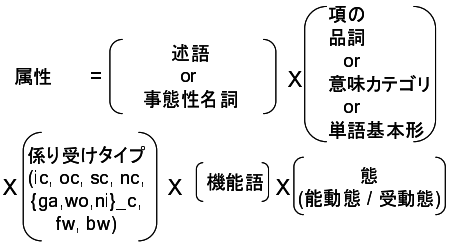


図 **1.** 属性 (= 制約 = 決定リストルール)

と述語（事態性名詞）が同一文内で係り受け関係にあるもの)、**文内ゼロ代名詞** (項と述語が同一文内にあるが係り受け関係に無いもの)、**文外ゼロ代名詞** (項が述語を含んでいる文とは別の文にあるもの (ただし同一記事内))、**同一文節内** (項と述語が同じ文節中にあるもの) の 4 つに分けた。なお外界照応に関しては、本稿では対象外とした。この表から述語については、「係り受け関係有り」のタイプの割合が高く、事態性名詞では、同一文節内についても比較的大きな割合を占めることが分かる。

3. 最近接単語の属性を利用した述語項構造解析

本節では我々の提案手法について述べる。提案手法では、解析対象の述語に対し正解となる項が様々な制約の下での最も近い位置にある名詞（名詞句）であるという仮説のもとに解析を行う。この「様々な制約」は、図 1 のように基本属性として、述語（事態性名詞）の基本形、係り受けタイプ、汎化レベル、機能語、態の 5 種類を設定し、基本属性の直積からなる属性として表した。

3.1. 係り受けタイプ

日本語においては、文節中の機能語および文節間の係り受け関係が項構造を決める上で重要である。我々は、図 2 のように、係り受けのタイプを項から述語に係る (ic)、述語から項に係る (oc)、述語と項が同一文節内にある (sc)、項から別の項に係る (ga_c, wo_c, ni_c)、係り受けがない（nc）に分け、さらに機能語や態に係らず単純に述語の前か後ろに項が存在することを示す補助タイプ (fw,bw) を定義した。

3.2. 汎化レベル

我々は対象となる項候補それぞれに対して単語基本形、意味カテゴリ、品詞の 3 つのレベルの属性値を使って制約を構成した。品詞に関しては、京都大学テキストコーパスでタグ付けされている品詞を用いた。意味カテゴリに関しては、各単語に人手で日本語語彙大系 [14] の意味カテゴリのうち、第 3 階層および第 4 階層の意味カテゴリを使用して意味カテゴリを付与した。

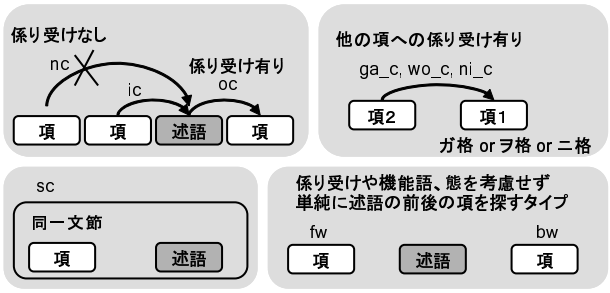
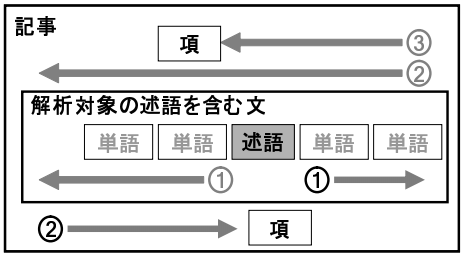


図 **2.** 係り受けタイプ



係り受けタイプ	単語サーチ方向
ic, bw, {ga, wo, ni}_c	←
oc, fw	→
nc, sc	1. ← 2. →

図 **3.** 訓練フェーズの概要

3.3. 機能語と態

我々は、項候補に後続する機能語と項候補の係り先用言の態（能動／受動）についても基本属性として用いた。

3.4. 訓練フェーズ

提案手法の訓練フェーズの概要について図 3 に示す。まず最初に、訓練データ中の対象述語に対して、正解となる項が様々な制約の下で最も近い位置である名詞句となるような属性を構成する。なお、近い位置を探す方向は図に示すように係り受けタイプにより異なる。次に、線形カーネルの Support Vector Machines (SVMs) [11] を用いて各項に対して有効な属性を one vs rest 手法で求めた。ここで、SVM の実装として TinySVM [1] を用いた。次に、線形 SVM が出力した重みをソートし、正の重みのルールのみを用いて決定リストを得る。最後に、各述語（事態性名詞）に対する項の出現確率を計算する。こ

[1] http://chasen.org/taku/software/TinySVM/

表 **3.** 述語に対する項存在確率の例

述語 または事態性名詞	項存在確率		
	ガ格 (主格)	ヲ格 (対格)	ニ格 (与格)
使う	44.7%	82.9%	5.3%
交渉	77.4%	30.7%	0.0%
参加	87.1%	0.0%	72.5%
基づく	81.9%	0.0%	100.0%

れは、項候補と述語の間に係り受け関係があるもので最終的に項と予測されたものが無かった場合に、述語と係り受け関係がない項候補に関しても決定リストを適用するかどうかを決定する際に使用される。

表 3 に、述語および事態性名詞に対する項存在確率の例を示す。事態性名詞「交渉」のニ格、事態性名詞「参加」のヲ格、述語「基づく」のニ格のように極端な値を取るときには、その項を決定すべきかそうでないか、高い信頼度で決定できる。しかし、述語「使う」のガ格のように確率が 50%に近いようなものは、項が実際に存在するかどうか決定するのは非常に難しい。

また、我々が使用した訓練データは、2874 述語に対し 6 万 2264 サンプルと、機械学習の面からは決して多いわけではなく、従来決定リストを学習するのに使われているエントロピーによる手法といったものでは精度が悪くなると考えられたため、汎化性能の高い SVM を決定リストの学習に用いた。

3.5. テストフェーズ

テストフェーズでは、訓練フェーズで学習された決定リストと項存在確率を用いてテストサンプルを解析する。まず、係り受けタイプにおいて、係り受けが存在するタイプ ic, oc, ga_c, wo_c, ni_c の属性を持つ決定リストを用いて、項の決定を行う (ステップ 1)。これは、日本語においては項構造を決定する際に、文法的な制約が意味的な制約よりも強いケースが多いためである。次に、ステップ 1 で決定されなかった項に対して、同一文節内タイプ sc の決定リストを用いて項構造の決定を行う (ステップ 2)。このステップは事態性名詞に対してのみ行う。これは、日本語の事態性名詞は項と複合語を形成して同一文節となるケースが多いためである。次に、項の存在確率を使って、次のステップに進むかどうかを決定する (ステップ 3)。項存在確率がある閾値（本稿では 50%）を下回るときには、ここでテストを終了する。最後に、係り受け関係がないタイプ、すなわち、nc, fw, bw タイプの属性を持つ決定リストを用いて項の決定を行う (ステップ 4)。このステップは、対象となる項が文法的には必要で、係り受け関係、機能語、態といった文法的な手がかりなしに、項と述語（事態性名詞）の間の共起関係だけで決定されるような場合のために実行される。

4. 実験結果

4.1. 実験データ

実験データとしては、NAIST テキストコーパス 1.4β [3] 全体を訓練用データ（記事：1 月 1 日〜1 月 11 日、社説：1 月〜8 月）、開発用データ（記事：1 月 12 日、1 月 13 日、社説：3 月）、テスト用データ（記事：1 月 14 日〜1 月 17 日、社説：10 月〜12 月）の 3 つの集合に分割し、訓練用データを用いて学習を行い、テスト用データを用いて精度評価を行った。なお、この訓練データ集合にはのべ 4 万 9527 個の述語と 1 万 2737 個の事態性名詞が含まれていた。

文節間の係り受け関係については同じ新聞記事を対象としている京都大学テキストコーパスにおける正解情報を用いた。また、項候補となる名詞句は複数単語、複数文節から構成されるものもあるが、訓練フェーズにおいて NAIST テキストコーパスでタグが付与された名詞句を用い、テストフェーズでは、簡単な人手ルールを用いて自動的に抽出した名詞句を用いた。

4.2. 比較手法

提案手法を評価するためのベースライン手法としては、述語/事態性名詞に係る名詞で、その所属文節における機能語が「が」「を」「に」であれば、それぞれその名詞を「ガ格」「ヲ格」「ニ格」の項とする手法を試みた。また、事態性名詞が複合語を形成する場合に関しては、もし複合語の中の事態性名詞以外の単語が、その事態性名詞と高い確率で訓練データの中で項として共起している場合、項と判定することとととした。

また、決定リストを作成するための従来手法としては、ルールによって選ばれたサンプルのエントロピーを用いて決定リストを作成する方法がある。我々は、このうち Yarowsky が提案した手法 [13] で決定リストを学習した実験も比較のため行った。

4.3. 実験結果

全体の実験結果を表 5 に示す。テストデータ全体を通して、提案手法の F 値による評価は、ベースライン手法および Yarowsky 手法を上回っている。

また、提案手法によって得られた決定リストのルール数は 1 述語あたり平均 103 個であった。表 4 に事態性名詞「交渉」に対する、ic タイプ（項が述語（事態性名詞）に係るタイプ）に関する決定リストを示す。ここで、「汎化レベル」の欄の「単語基本形」は、述語（事態性名詞）の基本形を表し、「意味カテゴリ」は、その意味カテゴリを表している。ic タイプの決定リストでは、＜

表 **4.** 事態性名詞「交渉」に対する ic タイプの決定リスト（上位 10 ルール）

順位	{ 係り受けタイプ, 汎化レベル, 意味的主辞, 機能語, 態 }	重み	出力格
1	{ic, 単語基本形 , 朝鮮民主主義人民共和国, の, 能動態 }	0.982	ガ格
2	{ic, 意味カテゴリ , ＜地域＞ , の , 能動態 }	0.638	ガ格
3	{ic, 単語基本形 , 日米両国 , の , 能動態 }	0.550	ガ格
4	{ic, 単語基本形 , 合弁会社設立 , の , 能動態 }	0.529	ヲ格
5	{ic, 単語基本形 , 電気通信分野 , の , 能動態 }	0.414	ヲ格
6	{ic, 単語基本形 , 朝鮮民主主義人民共和国, との, 能動態 }	0.317	ヲ格
7	{ic, 単語基本形 , 行為 , の , 能動態 }	0.308	ヲ格
8	{ic, 意味カテゴリ , ＜未分類語＞ , の , 能動態 }	0.294	ガ格
9	{ic, 単語基本形 , 自動車・同部品分野 , の , 能動態 }	0.278	ヲ格
10	{ic, 意味カテゴリ , ＜場＞ , の , 能動態 }	0.247	ヲ格

表 **5.** 実験結果（F 値（%））

述語	baseline	Yarowsky	提案手法
係り受け関係あり	72.7	79.4	**82.8**
同一文節内	0.0	59.2	**73.1**
文内ゼロ代名詞	0.0	9.1	**25.6**
文外ゼロ代名詞	2.2	9.3	**22.0**
事態性名詞	baseline	Yarowsky	提案手法
係り受け関係あり	19.6	56.5	**65.4**
同一文節内	49.0	61.5	**76.9**
文内ゼロ代名詞	0.0	16.9	**32.5**
文外ゼロ代名詞	1.3	16.0	**21.1**

地域＞＜場＞といった意味カテゴリの属性も上位のルールを占めている。

5. おわりに

本稿では日本語の述語項構造解析に関して、様々な制約の下で述語に対する最も近い単語の属性に基づく決定リストを用いた手法を提案した。本手法は、項に対する異なる属性の相対的な重みを学習し、その重みをソートして得られた決定リストを用いて項の決定を行う。提案手法を用いることで、項の決定の知識とゼロ代名詞解析の知識を統合し、日本語述語項構造解析において高い精度を実現できた。特に、事態性名詞に関してコーパスから様々なレベルの知識を抽出することができた。将来的には、より豊かな制約を用いた項構造解析の研究を行っていきたいと考えている。

謝辞

NAIST テキストコーパスの仕様および機能語の定義に関して貴重なコメントを頂きました、東京工業大学の飯田龍氏ならびに奈良先端科学技術大学院大学の松本裕治教授に感謝いたします。

参考文献

[1] Hirschman, L., Robinson, P., Ferro, L., Chinchor, N., Brown, E., Grishman, R. and Sundheim, B.: Hub-4 Event'99 General Guidelines (1999).

[2] Iida, R., Inui, K. and Matsumoto, Y.: Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution, *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 625–632 (2006).

[3] Iida, R., Komachi, M., Inui, K. and Matsumoto, Y.: Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations, *Proc. of ACL 2007 Workshop on Linguistic Annotation*, pp. 132–139 (2007).

[4] 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).

[5] Komachi, M., Iida, R., Inui, K. and Matsumoto, Y.: Learning-Based Argument Structure Analysis of Event-Nouns in Japanese, *Proc. of the Conference of the Pacific Association for Computational Linguistics (PACLING)*, pp. 120–128 (2007).

[6] Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A. and Popowich, F.: Description of SQUASH, the SFU question answering summary handler for the DUC-2005 summarization task, *Proc. of DUC 2005* (2005).

[7] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R.: The NomBank Project: An Interim Report, *Proc. of HLT-NAACL 2004 Workshop on Frontiers in Corpus Annotation* (2004).

[8] Narayanan, S. and Harabagiu, S.: Question answering based on semantic structures, *Proc. of the 20th International Conference on Computational Linguistics (COLING)* (2004).

[9] Palmer, M., Kingsbury, P. and Gildea, D.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106 (2005).

[10] Shen, D. and Lapata, M.: Using Semantic Roles to Improve Question Answering, *Proc. of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL)*, pp. 12–21 (2007).

[11] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer-Verlag, New York (1995).

[12] Walker, M., Iida, M. and Cote, S.: Japanese Discourse and the process of Centering, *Computational Linguistics*, Vol. 20, No. 2, pp. 193–233 (1994).

[13] Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 88–95 (1994).

[14] 池原　悟，宮崎正弘，白井　諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林　良彦：日本語語彙大系，岩波書店 (1997).

[15] 橋田浩一：GDA 日本語アノテーションマニュアル (2005). http://i-content.org/gda/tagman.html.

[16] 中岩浩巳：日英対訳コーパス中のゼロ代名詞とその指示対象の自動認定，情報処理学会研究報告（自然言語処理研究会）NL-98-1, pp. 33–40 (1998).