単語の意味分類と大規模日本語 n-gram を用いた統計的日本語述語項構造解析

若木 裕美　　市村 由美　　住田 一男　　永江 尚義
(株) 東芝 研究開発センター 知識メディアラボラトリー
{hiromi.wakaki, yumi.ichimura, kazuo.sumita, hisayoshi.nagae}@toshiba.co.jp

1 はじめに

述語項構造解析用のコーパスとして一般に用いられる NAIST テキストコーパスは新聞記事にアノテーションされているため、既存研究では新聞記事を元に作成された知識や比較的少量の Web データに対する精緻化された情報を利用することが多い。例えば、今村ら [1] では 12 年分の毎日新聞コーパスを、松林ら [2] は京大格フレームと文脈類義語データベースを用いている。今後さまざまな応用に広く利用するためにはドメインに依存しない、より一般性の高い知識や述語項構造の特徴を汎化しやすい素性を用いるのが望ましいと考える。そこで、単語の意味に汎化した素性として固有表現と一般名詞の両方に意味分類を与えて素性に利用した。単語の意味分類は我々の開発した解析器で判定した。また、幅広い語彙に対応するため大規模日本語 n-gram から獲得した統計情報も素性に利用した。本稿では、実験により NAIST テキストコーパスに対してこれらの素性の有効性を確認する。

2 解析手法

本研究では新設した単語の意味分類を用いた素性と大規模日本語 n-gram を用いた素性の有効性を確認することを主眼に、基本的なタスク設定、解析手順は松林ら [2] と同一にし性能比較する。文内の項のみを解析対象として、与えられた述語に対する項候補の格を推定するモデルを作成した。松林らと同様に最右形態素を出力とし、形態素単位で評価を行うこととした。

2.1 解析モデル

解析モデルは松林らに類似しているが、細かい点で異なるため本節で説明する。

1. 項候補の選択
   NAIST テキストコーパスの品詞分析を行い、正解のカバー率が高くまた正解確率が高い品詞を選別した。その結果、名詞（16 種類の細品詞）[1]、記号、アルファベット、助詞の「副助詞／並立助詞／終助詞」、連体詞を項候補の品詞とした。品詞選択によりカバー率 97.5% で、38.1% に圧縮できた。さらに項候補は、係り有・文内係り無・文節内を想定し述語と係り受け有の文節の主辞、係り受け無で述語より前に出現する同一文内の文節の主辞、述語と同一文節内の全形態素を対象とした。係り有と文内係り無ではそれぞれ主辞だけでカバー率 98.1%、93.9% だったが、文節内では 25.7% しかなかったためである。なお、動詞については事例分析の結果から、次の形態素が同一文節の末尾でありその品詞が「助詞, 格助詞-一般」だけをさらに項候補に利用した。

2. 格推定
   与えられた述語と前項により選択された項候補についてガ、ヲ、二、NONE の 4 種類への多クラス分類として最大エントロピー法により学習し[2]、項候補について各ラベルに対する確率を求める。なお NONE は述語項関係にない場合を表す。1 つの述語に対しガ、ヲ、二の各ラベルについて全項候補の中で最も確率の高い項候補をそれぞれ 1 つ選択する。また、閾値を定め、設定した閾値を超えた確率の項候補のみを対象述語の項とし、格ラベルとともに出力する。

2.2 素性

我々の解析で利用した素性を表 1 に列挙した。存研究に使われている基本的な素性に倣って類似の素性設計を行ったが、異なる点について説明する。また、本研究で新設した意味クラスに基づく素性と大規模日本語 n-gram を用いた素性について説明する。

述語と項の素性　述語/事態性名詞と項は、その形態素位置に基づき素性を生成するが NAIST テキストコーパスのアノテーションの基準に基づき意味的に同一視できるものを共通化するため次のようにした。まず、与えた述語形態素と項候補の形態素はラベル判定する

[1] 固有名詞-組織、接尾-助数詞、代名詞-一般、副詞可能、接尾-人名、固有名詞-地域-一般、形容動詞語幹、固有名詞-人名-姓、ナイ形容詞語幹、接尾-地域、固有名詞-一般、固有名詞-人名-名、接尾-特殊、非自立-副詞可能、接尾-サ変接続、固有名詞-人名-一般の 16 種。

[2] 学習には、岡崎氏が開発した Classias を利用した。

[3] 各格 c には、「が、を、に、へ、で、まで、から」を使った

表 1: 利用した素性一覧。(*) で示した素性は複数個の素性になるもの。素性 28〜30 の式で用いた記号の説明は 2.2 節を参照。

	ID	素性の説明
述語に関する素性	1	述語形態素の原形
	2	述語形態素の読み
	3	述語形態素の品詞と細品詞の連結
	4	述語の主辞以降の各単語 (*)
	5	「れる/られる」「せる/させる」「できる/出来る」の有無の連結
項候補に関する素性	6	項形態素の原形
	7	項形態素の読み
	8	項形態素の品詞と細品詞の連結
	9	項の機能語
	10	項の助詞部の各単語 (*)
	11	項の助詞部の文字列
	12	項の主辞までの文字列
述語と項候補に関する素性	13	述語と項の文節の係り受けの有無と係り受け有の場合はその向き
	14	素性 1, 5, 13 の組み合わせ
	15	素性 1, 5, 10, 13 に述語の文節の各単語を組み合わせたもの (*)
	16	素性 5, 10, 13 に述語の文節の各単語を組み合わせたもの (*)
	17	素性 1, 6 の組み合わせ
	18	素性 5, 11, 13 の組み合わせ
	19	素性 1, 8, 13 の組み合わせ
	20	素性 1, 5, 8, 11, 13 の組み合わせ
	21	素性 3, 8, 13 の組み合わせ
	22	素性 3, 5, 8, 11, 13 の組み合わせ
述語と項候補の間に関する素性	23	素性 5 と述語と項の間にある助詞をつなげたもの
	24	素性 5 と 13 と述語と項の間にある助詞をつなげたもの
	25	述語にかかる文節の主辞以降の文字列
	26	係り有で述語が項候補にかかる場合に述語の直前の文節の助詞だけと、素性 1 とその助詞の組み合わせ (*)
	27	同じ文節か
n-gram に関する素性	28	$log_{10}(f(c)+1)$ を格助詞 c ごとの素性の値として発火[3](*)
	29	c が { が、を、に } のうち最も頻度 $f(c)$ の高いラベル c_{max} に対してのみ発火し $\frac{f(c_{max})}{\Sigma_{c_i=\{\text{ が, を, に }\}} f(c_i)}$ を素性の値とする
	30	$log_{10}(f_{class}(c)+1)$ を格助詞 c ごとの素性の値として発火[3](*)

位置とし、基本的にはこの形態素から述語や項に関する素性を生成する。ただし、ラベル判定用の述語形態素原形が「する」でその 1 つ前の形態素品詞が名詞である場合には、述語の素性を生成する形態素として 1 つ前の形態素を使い、表 1 の素性 1〜3 を生成した。

また、項の助詞部の文字列 (素性 11) には、基本的に項文字列より後ろの形態素列を取るが、項が述語より前にある同一文節内の場合のみ項と述語の間の形態素とした。項の助詞部の各単語 (素性 10) には素性 11 の形態素を使った。

大規模日本語 n-gram を用いた素性（素性 28, 29） Google n-gram は Google が Web から抽出した約 200 億文（約 2550 億単語）の日本語データから作成した n-gram データである。対象文数が多くまた Web データであり広範な分野を包含していると考えられるため、本実験では格フレーム辞書の代わりに Google n-gram の 3-gram の 1 語目を項候補、2 語目を格助詞 c、3 語目を述語と見立ててその頻度を素性に利用した。項候補と述語に対する格助詞 c ごとの頻度 $f(c)$ を「項候補＋{ が、を、に、へ、で、まで、から }＋述語」の 3-gram 頻度により得た。素性 28 では、$log_{10}(f(c)+1)$ を格助詞 c ごとの素性の値とした。さらに、素性 29 では、{ が、を、に } の中で最も頻度 $f(c)$ の高い格助詞 c_{max} に対してのみ発火させ、$\frac{f(c_{max})}{\Sigma_{c_i=\{\text{ が, を, に }\}} f(c_i)}$ を素性の値とした。

単語の意味分類で汎化した素性（素性 30） Google n-gram の 3-gram に対し単語の意味分類で汎化した 3-gram を作り、その頻度 $f_{class}(c)$ を素性に利用した。

まず、単語の意味分類には、我々の開発した意味クラス解析器が出力する意味クラス [3] を用いる。意味クラス解析は固有表現に限らず一般名詞にもラベルを付与することが特徴で、意味クラスには固有表現抽出で一般に対象とされる分類『人名』『地名』『組織名』『日時』のほか『イベント』『交通機関』『商品』『食べ物』『動物』などからなる。例えば、「議員」や「お母さん」などにも『人名』ラベルを付与する。「運動会」のような一般名詞でも「○○小学校第１２回運動会」のような固有表現も共に『イベント』ラベルを付与する。

次に、3-gram を意味クラスで汎化する方法について述べる。意味クラス解析器の出力は、曖昧性のある入力に対して複数の意味クラスを出力する。例えば「川崎」という語は、『人名』と『地名』の２つの意味クラスが付与される。そこで、格助詞ごとの述語に対する意味クラスの重みを計算しておき、それを元に 3-gram を意味クラスで汎化した頻度を生成する。ここで、各 3-gram を g_j とし、その 3 語を前から順に w_1〜w_3 と

記述することにし、次のように計算する。

1. w_2 が格助詞 c で w_3 が述語 $pred$ であるときの w_1 を意味クラス解析した結果の意味クラス群 $Class(g_j)$ を元に、格助詞ごとの述語に対する意味クラスの重み $W_c(class_i, pred) = \Sigma^{全\ g_j}_{class_i \subset Class(g_j)} \frac{1}{|Class(g_j)|}$ を算出。
2. $f_{class}(class_i, c, pred) = \Sigma_{全\ g_j} \Sigma_{class_i \subset Class(g_j)}\ f(g_j) \times W_c(class_i, pred)$ を合計し 3-gram を再構成した。

例えば、3-gram「川崎/に/行く」では意味クラス付 3-gram として「『地名』/に/行く」と「『人名』/に/行く」の 2 つが得られる。ステップ 1 ではそれぞれ $\frac{1}{2}$ ずつの重みとし、すべての 3-gram で計算した重みを合計しておく。ステップ 2 では、3-gram の頻度とステップ 1 の重みを掛けた頻度を合計する。

そして、素性 30 の生成には項文字列を入力として項候補の意味クラスを得る。項文字列は、基本的に項の文節の先頭から主辞までの形態素列とする[4]。一方、述語が同一文節内で項より前にある場合は項形態素のみとする。そして、項候補の意味クラス解析結果の中で $W(c, pred, class_i)$ を最大にする意味クラス $class_{max}$ と述語 $pred$ から算出できる各格助詞 c ごとの頻度 $f_{class}(class_{max}, c, pred)$ を $f_{class}(c)$ として、$log_{10}(f_{class}(c)+1)$ を格助詞 c ごとの素性の値とした。

3 実験

本実験では、NAIST テキストコーパス (NTC) をコーパスとして、意味クラス素性や n-gram 素性の素性の貢献度を確認する。

コーパス　NTC を学習・開発・評価データとして利用する。NTC には NTC1.4b 版と NTC1.5 版の 2 バージョンあるが、本稿で主要な比較対象とする松林ら [2] に合わせて開発には NTC1.5 版を利用した。一方、既存研究の多くは NTC1.4b 版を利用しているため、比較のため NTC1.4b 版での評価も行った。

学習・開発・評価データの分割は、多くの既存研究と同様に 1 月 1〜11 日のニュース記事と 1〜8 月の社説記事を学習データ、1 月 12〜13 日のニュース記事と 9 月の社説記事を開発データ、1 月 14〜17 日のニュース記事と 10〜12 月の社説記事を評価データとした。

NTC1.5 と NTC1.4b では形式が異なるため、次のように変換した。まず NTC1.5 は京都テキストコーパス 4.0 に対応し、各形態素にラベルが付与されている。そこで、NTC1.4b の XML 形式から NTC1.5 と同様に最右形態素にラベルを付与しなおした。NTC1.4b は京都テキストコーパス 3.0 を元にしており形態素単位が異なる事例があったため、NTC1.4b の XML で項とされた範囲の最右文字を含む形態素に割り当てた[5]。

[4] 項形態素だけを入力にすると正しく解析できない場合があるためである。例えば「五十人が」という文節の場合、「五/十/人/が」という形態素列に対し、主辞は「人」、項形態素は「人」である。意味クラス解析は、「五十人」を入力とすると意味クラス『人数』を出力するが、「人」に対しては『人名』出力する。

[5] 松林らは京都テキストコーパス 3.0 に合わせている。

表 2: 事例数

		述語			事態性名詞		
		訓練	開発	評価	訓練	開発	評価
係り有	ガ	37666	7547	14248	3928	827	1564
	ヲ	24997	5107	9532	2531	544	1052
	ニ	5854	1637	2547	279	74	102
文節内	ガ	168	52	76	3960	1193	1879
	ヲ	124	33	57	5170	1326	2171
	ニ	331	101	132	508	122	201
係り無	ガ	11552	2554	4768	7344	1897	3159
(文内)	ヲ	1802	394	783	1384	387	632
	ニ	359	112	211	542	153	215
合計		82853	17537	32354	25646	6523	10975

(a) NTC1.5

		述語			事態性名詞		
		訓練	開発	評価	訓練	開発	評価
係り有	ガ	36445	7508	13949	2925	592	1159
	ヲ	24012	4990	9290	1882	359	795
	ニ	11042	2859	5403	159	55	113
文節内	ガ	162	31	30	1305	322	555
	ヲ	83	9	19	3322	777	1433
	ニ	476	32	97	430	205	201
係り無	ガ	11163	2531	4637	5031	1366	2384
(文内)	ヲ	1796	406	761	829	221	423
	ニ	455	171	300	271	100	181
合計		85634	18537	34486	16154	3997	7244

(b) NTC1.4b

評価には松林らと同様に共参照関係にある形態素のいずれかと一致すれば正解したとしてカウントした。しかし、NTC1.4b では形態素ごとに述語項の ID が異なるため述語に対して唯一の項が明らかなのに対し、NTC1.5 では共参照関係にある項にすべて同じ述語項 ID が与えられているため、評価時の係り有等の分類が自明でない[6]。そこで、今回は共参照関係にある項のうち文節内、係り有、文内係り無の順に存在した分類を評価用の分類とした。表 2 にこの分類結果に基づいた事例数を示した。松林らの文献に記載された NTC1.5 の事例数と比較したところ、評価データと開発データの述語合計、事態性名詞合計が一致しており比較可能な値といえる[7]。なお、学習用の正解ラベルには、共参照関係にある項すべてに正解ラベルを与えて学習した。

結果と考察　表 3 と表 4 に既存研究の松林ら [2] との F 値比較を示した[8]。なお、評価には共参照関係にある形態素のいずれかと一致すれば正解にカウントしたが、NTC1.4b では述語に対し唯一の正解の項がアノ

[6] 松林らの論文には基準の記載はなかった

[7] 文節内では同数だったが、係り有と文内係り無にカウントした数が互いに同数増減していた。学習データは 3 事例だけ異なっていた。NTC1.4b の事例数は記載が無かった

[8] 松林らが事例数を公開していた NTC1.5 に関し、表 4 と表 5 に関する事例数は本実験と一致するため比較可能な値と言える。一方、表 3 は文節内以外の分類では事例数が一致していないため厳密には比較できない。

表 3: NTC における F 値比較の詳細。(NTC1.4b の結果では、(A)NTC1.5 と同様の評価方法で共参照の項をすべて正解とした場合、(B) 共参照を考慮しない場合。)

			係り有				文節内				文内係り無				All
			ガ	ヲ	ニ	All	ガ	ヲ	ニ	All	ガ	ヲ	ニ	All	All
NTC1.5	述語	松林ら	87.8	94.0	63.7	87.1	37.3	43.2	65.2	54.5	49.0	27.7	25.7	**45.8**	80.9
		提案手法	90.3	95.0	61.7	**89.0**	39.7	63.2	63.5	**58.1**	41.8	26.0	9.1	38.7	**81.5**
	事態性名詞	松林ら	69.7	74.6	43.5	70.3	71.2	82.1	54.3	**76.2**	45.2	25.0	23.3	**41.6**	**63.5**
		提案手法	74.1	76.6	47.7	**73.9**	66.7	83.8	59.2	75.5	41.2	29.9	23.1	38.6	63.2
NTC1.4b (A)	述語	松林ら	85.1	92.7	83.1	87.2	23.5	18.5	63.2	**53.2**	47.1	30.3	23.3	**44.1**	81.0
		提案手法	88.4	94.3	83.0	**89.4**	13.0	34.9	60.5	50.8	41.8	29.2	6.6	38.6	**82.3**
	事態性名詞	松林ら	74.0	74.4	57.1	73.3	71.1	86.4	47.5	79.5	47.4	25.4	28.5	**43.9**	64.8
		提案手法	77.4	77.8	57.3	**76.6**	76.0	88.0	49.1	**81.9**	43.6	32.7	22.5	41.0	**65.5**
NTC1.4b (B)	述語	提案手法	88.4	94.3	83.0	89.3	12.8	34.9	60.5	50.6	41.4	29.0	6.1	38.2	82.3
	事態性名詞	提案手法	77.3	77.7	57.3	76.5	75.6	88.0	49.1	81.8	43.0	32.4	22.5	40.5	65.3

テーションされているため、共参照を考慮しない結果も載せた。それぞれ表中の (A)(B) で示した。

表 4 の NTC1.5 の全体性能では 77.1 で本手法が若干上回った。表 3 の最右列の結果を見ると、NTC1.5 では述語と事態性名詞のそれぞれ 81.5、63.2、NTC1.4b ではそれぞれ 82.3、65.5 で、NTC1.5 の事態性名詞を除いていずれも本手法が上回った。述語項の関係ごとの分類における全ての格の結果 (All) の比較では、係り有ではすべて本手法が上回った。一方、節内では互いに結果が良い場合と悪い場合があり、文内係り無ではすべて松林らの結果が良かった。本手法の素性には文内係り無に関する素性が不足していたと考えられる。

次に、表 5 に開発データでの文内の項に対する F 値を基準に、素性 28〜30 を除いた際の F 値を示した。素性 28〜30 をそれぞれ 1 つずつ除いた場合には、素性 28 が 76.75 で基準の 77.31 から一番大きく下がっており、最も効果が高いといえる。一方、素性 28〜30 のいずれの組み合わせで除外しても基準 77.31 を超えることはなく、素性 28〜30 をまとめて利用するのがもっとも良いことが分かった。

さらに、本実験を通じて分かったコーパスの分析結果を述べる。事例数の分析 (表 2) を見ると述語の係り有のニ格が NTC1.5 は NTC1.4b に比べて約半減していた。一方、F 値 (表 3) の述語の係り有では、松林らと本手法でそれぞれ、NTC1.4b で 83.1、83.0 に対し、NTC1.5 では 63.7、61.7 で、両手法で F 値が約 20 下がっていた。失敗分析の結果、NTC1.4b に付与されていたニ格が NTC1.5 にない事例が散見された。例えば NTC1.5 で「ブラジルに渡る」に対し学習器はニ格を出すが、正解にはニ格がなく不正解となっていた。このことからニ格のアノテーションには NTC1.4b と NTC1.5 で違いが大きく、コーパス利用時に考慮が必要な場合があることが分かった。

表 4: NTC における F 値比較

		NTC1.5	NTC1.4b
ALL	松林ら	76.8	未記載
	提案手法	**77.1**	79.5

表 5: NTC1.5 開発データでの文内の項を対象とした F 値比較による素性の貢献度。All は表 1 の全素性を表し、– で除外した素性を示した。)

	素性パターン	F 値	基準との差
(基準)	All	**77.31**	
	All–{ 素性 28,29,30}	76.01	-1.30
	All–{ 素性 29,30}	77.27	-0.04
	All–{ 素性 28,30}	76.84	-0.47
	All–{ 素性 28,29}	75.92	-1.39
	All–{ 素性 28}	76.75	-0.56
	All–{ 素性 29}	77.28	-0.03
	All–{ 素性 30}	77.25	-0.06

4 おわりに

本実験で、大規模日本語 n-gram を使った素性、意味クラスで汎化した n-gram 素性を利用し、NAIST テキストコーパスにおいてその有効性を確認した。今後は対話コーパスでの検証を行う。

参考文献

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

[2] 松林優一郎, 乾健太郎. 統計的日本語述語項構造解析のための素性設計再考. 言語処理学会第 20 回年次大会, pp. 360–363, 2014.

[3] 石谷康人, 鈴木優, 布目光生. 意味クラス解析と意図推定に基づくインタラクティブな情報検索インタフェース. 情報処理学会論文誌, Vol. 48, No. 12, pp. 3793–3808, dec 2007.