

# 事態性名詞の項構造解析における共起尺度と 構文パターンの有効性の分析

小町守, 飯田龍, 乾健太郎, 松本裕治  
奈良先端科学技術大学院大学 情報科学研究科  
{mamoru-k,ryu-i,inui,matsu}@is.naist.jp

## 1 はじめに

我々は「推薦」のような事態を含意する名詞を事態性名詞と呼び, たとえば「彼は上司の推薦で抜擢された」という文から「上司が彼ヲ推薦(する)」といった関係を抽出することを目標にしている. 事態とは行為や状態, 出来事を指し, 述語項構造と同様の項構造を考えることができる. 述語項構造を正しく認識することは, 情報抽出や自動要約, 機械翻訳といった応用で重要である.

述語項構造解析では, Gildea ら [1] が機械学習を用いた解析手法を提案した. また, 近年言語資源としても Penn Treebank に対して事態性名詞の項構造を付与したコーパス NomBank[2] が作成され, 日本語においても事態性名詞の項を含む名詞間の関係を付与した京都テキストコーパス第 4.0 版 [3] が公開された. 我々も NAIST テキストコーパス [4] に述語項構造を付与し, 事態性名詞の項構造解析を行う基盤が整ってきた.

事態性名詞の項構造解析を述語項構造解析と比較すると, 2 つの大きな違いがある. 1 つ目は名詞の多義性の問題であり, もう 1 つは解析単位の問題である.

1 つ目の多義性の問題とは, 事態性名詞の中には文脈によって事態を指す用法と指さない用法といずれも持つ名詞があるため, 曖昧性を解消しなければならない問題である. 事態性名詞とは事態を指す用法を持つ名詞のクラスを指し, 事態を指す用法で使われているとき事態性があると定義する. サ変名詞に限定すると, 事態性のある場合は動詞としての用法を考えたときの項を基本的に受け継ぐ<sup>1</sup>ため, これらの項を同定したい. たとえば「公衆電話で電話をすることがめっきり減った」という文では, 電話はモノとしての電話であり, これ自体はなんらかの事態を指す用法ではないが, 電話は「電話をする」という行為であり, 「(話し手)ガ(誰か)ニ電話をする」という事態を指している.

そこで, 我々は事態性名詞の項構造解析を大きく 事

態性判別と 項同定に分ける. 事態性判別は事態性名詞が事態を指す用法として使われているかどうか判別するタスクで, 項同定は事態を指す事態性名詞の項がどれかを同定するタスクである.

前回の予備調査 [5] では同一文節内に項が頻出するヲ格を対象に試験的に解析したが, 今回は解析の対象を限定せず, 文内の項同定タスクに有効な情報を検討する.

2 つ目の解析単位の問題とは, 項同定の単位の問題である. 日本語の述語項構造解析では, 項は「民間支援」が活性化するように述語「活性化する」の項となっているガ格の「民間支援」は文節単位で考えればよいが, 事態性名詞の項構造解析においては事態性名詞「支援」の項のヲ格が同一文節内の主辞以外の形態素「民間」を指しているように, 形態素単位で考える必要がある.

実際の NAIST テキストコーパスでの事態性名詞と項の出現位置の分布をまとめた表 1 を示す. 文内<sup>2</sup>・文間<sup>3</sup>・外界<sup>4</sup>の総数と, 文内に関しては事態性名詞と同一文節内に項がある場合, および事態性名詞が項に係っている場合と項が事態性名詞に係っている場合の数を記した.

表 1 から分かるように, ヲ格の場合約半数が同一文節内の主辞以外に項が来るなど, 文節内の語構成情報も項同定に用いる必要がある. また述語と比較すると文間に項が現れる割合が 1.5 倍程度多く [4], 構造上の手がかりに重点を置いて項を同定することが難しい.

構造以外の情報も用いた項同定に向けて, 以下 2 節で動詞と格要素の共起を用いて「項らしさ」を判定する方法を検討し, 3 節で共起によって解くことができない問題に対して事態性名詞に特有の構文パターンを調査する. さらに, 4 節で共起と構文パターンのいずれを用いても解くことができない問題についても考察する.

<sup>1</sup>「動き」「薦め」などの和語動詞由来の事態性名詞も同様に扱うことができる. また, 「運動会」のように事態を指す動詞と直接の結びつきがない事態性名詞は今回扱わない.

<sup>2</sup>事態性名詞と同じ文中に項が出現

<sup>3</sup>同一文書内で異なる文に項が出現

<sup>4</sup>文書内に項が存在しない

表 1: 事態タグと項の出現位置の分布

	文内			総数	文間 総数	外界 総数
	同一 文節	係り先				
		事態	項			
ガ	2024 (13%)	5220 (33%)	1208 (8%)	15583	5185	7501
ヲ	5519 (55%)	3127 (31%)	297 (3%)	10045	854	41
ニ	841 (49%)	352 (20%)	66 (4%)	1729	201	10

## 2 動詞と格要素の共起尺度の有用性

まず動詞と格要素の共起モデルが事態性名詞の項同定に利用できるかどうかを確認するため、事態性名詞をサ変名詞に限定して動詞と格要素の共起モデルを調査した。

動詞と格要素の共起モデルには藤田ら [6] が提案したモデルがある。藤田らは名詞  $n$  が格助詞  $c$  を介して動詞  $v$  に係っているときの共起確率  $P(\langle v, c, n \rangle)$  の推定に PLSI [7] を用い、 $\langle v, c, n \rangle$  を  $\langle v, c \rangle$  と  $n$  の共起と見なしてモデルを作成した。

$$P(\langle v, c, n \rangle) = \sum_{z \in Z} P(\langle v, c \rangle | z) P(n | z) P(z)$$

$Z$  は共起に関する潜在的な意味クラスを指す確率変数で、確率分布を用いて単語行列を  $|Z|$  次元に圧縮していることに相当する。 $P(\langle v, c \rangle | z)$ ,  $P(n | z)$ ,  $P(z)$  は EM アルゴリズムによって推定する。

共起尺度の利用に関しては、文内に現れる他の名詞と比較した項らしさを判定するためには、他の名詞と比較した情報を用いることが必要であると考え、項となる名詞とそれ以外の名詞をペアにして比較する実験を行った。名詞のペアを入力としていずれが項らしいか判定する分類器は、文内の全名詞を対象にした総当たり戦でもっとも勝ち数が多かった名詞が項だと判定する場合でも、勝ち抜き戦で勝ち上がった名詞が項だと判定する場合でも共通して用いることができる。また、どのペアでも項を正しく認識できる分類器を作成することは、いずれの場合でも重要である。

実験には NAIST テキストコーパスから新聞記事 1 日分 137 記事 (1,226 文) 中、文内にガ格の項がある事態性名詞を対象に、項の名詞と、文内にある他の名詞を比較し、項であるか否かと共起スコア<sup>5</sup>の相関を図 1 にプロットした。<sup>6</sup>図の  $f(x) = x$  より上部の点は項のほうが

<sup>5</sup>共起確率を用いて相互情報量、 $\chi^2$  係数、Dice 係数をそれぞれ計算して比較したところ、相互情報量を用いた結果がもっとも相関がはつきり現れたので、相互情報量を用いた結果のみを示す。

<sup>6</sup>固有名詞を中心に共起モデルの作成時に出現しなかったデータが全体の 18% あったため、共起スコアを計算することができなかった名詞と共起スコアが負であった名詞に関して、CaboCha<sup>7</sup> の出力する固有表現ラベルおよび ChaSen<sup>8</sup> で品詞が固有名詞であると判定された名詞に対しては固有表現クラスで作成した共起を用いた。

共起スコアが高い事例、下部の点は他の名詞のほうが共起スコアが高くなってしまっている事例である。

この図の第二象限と第四象限に含まれる項と他の名詞のペアの総数は全体の 71.2% (9,715 事例) あり、共起スコアの大きい方が正解とした場合の精度は 90.0% である。つまり、第二象限と第四象限に限ると共起スコアだけでも高精度に項の判別が可能である。

第一象限の事例は 28.1% (3839 事例) あり、同様に精度を求めると 55.8% であり、共起スコアのみで判定することはできない。第三象限の事例は項と候補のペアのいずれも項らしくない事例だが、これらの事例には共起スコアの作成のときデータの過疎性によりスコアが低くなっている事例が含まれるので、名詞をクラスでまとめあげて共起スコアを計算すれば一部は残りの象限にマップすることができる。それでもなお残る事例については第一象限と第三象限ともに共起以外の情報を用いた判別を考える必要がある。

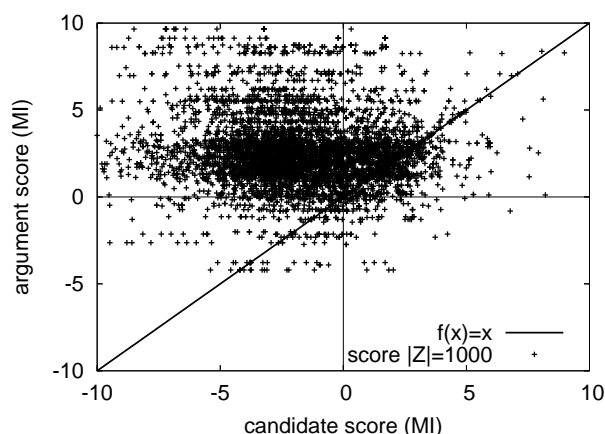


図 1: 項と項候補の共起スコアの比較

また、新聞記事約 30 年分から推定した共起モデルと、共起の取得対象をさらに大規模化した共起モデルとの比較を行った。共起の取得元となる大規模コーパスには河原ら [8] の作成した Web 5 億文コーパスを用いた。評価は新聞と同様、項と文内の他の名詞句とを比較し、共起スコアが大きい方が項であるとき正解、項でないときを不正解とした。

PLSI の隠れクラス数  $|Z|$  の最適な値は明らかではないので、いくつかの値を評価して図 2 に示した。Web の共起モデルでは隠れクラス数 4,000 のときが精度 87.2% (11,894/13,640) ともっとも高かった。新聞の共起モデルをこの問題に適用したところ、精度 90.4% (12,335/13,640) であった。Web の共起モデルのほうが精度が悪いが、これは評価に新聞記事を用いているの

<sup>7</sup><http://chasen.org/~taku/Software/CaboCha/>

<sup>8</sup><http://chasen.naist.jp/hiki/ChaSen/>

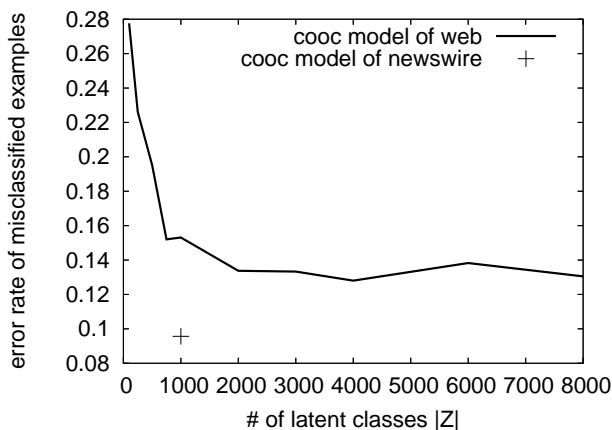


図 2: PLSI の  $|Z|$  を変化させたときの分類性能

で、共起の分布が異なるためだと考えられ、共起モデルの構築に利用する資源の選択については今後さらに検討を続けたい。

### 3 述語項構造解析モデルを利用した項同定実験

前節で述べたように、項の候補ペアがともに第一象限にある場合、つまりともに正の相関がある場合は、共起以外の情報を用いる必要がある。そこで、まず既存の述語項構造解析 [9] の項同定モデルで利用されている情報を使って文内にガ格・ヲ格・ニ格を持つ事態性名詞を対象に項と文内の他の名詞をペアで見たとときの分類を行い、それでも解けない第一象限の問題を人手で分析した。

実験には NAIST テキストコーパス 1.2 $\beta^9$ を用いた。2名の作業者によってタグ付与した新聞記事 2 日分のうち、両者で判定が一致した部分、新聞記事 1 日分 (137 記事) を訓練事例、それとは異なる 1 日分 (150 記事) を評価事例に用いた。実験に用いた機械学習器は Support Vector Machines [10] であり、実装は TinySVM<sup>10</sup>、多項式 2 次カーネルでパラメータはデフォルト値を使用した。

使用した素性は Iida らのモデル [9] に準じ、事態性名詞で計算することのできない素性<sup>11</sup>は使用しなかった。事態性名詞の項同定は形態素単位で行うため、文節内の位置の素性、同一文節内の他の形態素の品詞列の素性を追加した。また、村木 [11] に掲載されている機能動詞 128 表現を用い、事態性名詞が機能動詞に係っているかどうかの素性と、機能動詞に係っている場合に機能動詞と項に係り関係にあるかどうかの素性を追加した。

文内に項を持つ事態性名詞のみを使用し、正解の項と文内に出現する他の全ての名詞句を訓練事例にし、与えられた形態素対のどちらが項らしいかを判定するモデルを作成した。評価では、各項を文内の他の全ての名詞と

<sup>9</sup><http://cl.naist.jp/nldata/corpus/>

<sup>10</sup><http://chasen.org/~taku/Software/TinySVM/>

<sup>11</sup>事態性名詞にはヴォイスと助動詞は定めることができない。また、センタリング理論に基づいたヒューリスティック素性は用いていない。

表 2: 実験結果

素性	誤り個数
全ての素性を用いた場合	476 (3.5%)
- 語彙素性	510 (3.7%)
- Web 共起素性	502 (3.7%)
- 共起素性	684 (5.0%)
- 意味素性	510 (3.7%)
- 構文素性	644 (4.7%)
- 位置素性	700 (5.1%)
新聞共起のみを用いる	1,305 (9.6%)

表 3: 構文パターンに起因する誤り事例の数の誤りのタイプ

誤りのタイプ	事例数
機能動詞など動詞と項を共有する	19
項に係るときに格交替がある	13
複合名詞内に項がある	12
項が連体節に出てくる	6
項が A の B の形で出てくる	6
依存関係にあるが遠く離れている	6

組にしていずれが項であるか判定させた。全ての素性を用いた場合の結果と、それぞれの素性を全ての素性から取り除いた結果を表 2 にまとめた。また、比較のために共起スコアの低い方が項であるとした場合の結果も載せた。

実験の結果、事態性名詞の項同定にいちばん効果が高いのは位置素性、特に文節内のどの場所に出現したかという素性であることが分かった。また、次に効果が高いのは共起素性であり、項同定に共起情報の効果が高いことが分かった。構文素性の効果は位置素性や共起素性に次ぐ効果であった。

### 4 議論

間違えた事例のうち、新聞の共起モデルのスコアが第一象限にあり、なおかつ対立候補のほうが共起スコアが高い事例 112 個を分析したところ、以下のような傾向が見られた。

構文パターンを正しく扱っていない (62 事例)

分類を誤った事例のうち最も多かったのが、事態性名詞に特徴的な構文パターンをうまく使えていない事例である。典型的には以下のような例がある。

ドイツは……エリツィン政権に対する支援を続けるべきかどうか苦しい **選択** に迫られている。

ここに出てくる「X が Y (事態性名詞) に迫られる」というパターンでは、事態性名詞 **選択** のガ格は述語「迫られる」のガ格に一致する。このようなふるまいをする動詞は「X が Y を展開する」「X が Y を延期する」「X が Z を Y を要請する」「X が Y を踏み切る」などがあり、これらは係り先の述語と項を共有する。

今回の実験では [11] に挙げられている機能動詞を全て追加したが、上記に挙げた「展開する」などの例はカバーされていなかった。これらの動詞は主に事態性名詞のAspectやVoice, モダリティを変えるために使われるものであり、辞書的に書き尽くすことができると予想されるが、これらのパターンをいかに用いるかが重要であり、今後このような構文パターンの使い方を検討する予定である。

また、次に多かった事例は係り関係にある項の間違いである。「ロシアのチェチェン共和国に対する武力介入」などの事例では「の → ガ」「に対する → ニ」といった格の交替が起こっており、項同定に失敗している。同様な事例として「から → ガ」「について → ヲ」といったパターンが存在する。これらの事例についても明示的に格交替が起きるパターンを組み込んでいきたい。

残りの事例については複合名詞内の依存構造やAのB, 連体節の解析など、事態性名詞周辺の構文パターンを適切に捉える必要があり、それぞれどのようなパターンを抽出すればよいかは今後の課題である。62 事例の内訳は表 3 にまとめた。

#### 共参照を正しく解決できない (13 事例)

今回の項同定実験では共参照の解決は行わなかったが、  
.....マルチメディア市場を「市場規模百二十三兆円、雇用創出が二百四十万人」と予測したが.....

のような事例では、述べられている「市場」は実体ではなく名詞のクラスを指しているので市場と市場のいずれかを当てれば正解と見なすことができるが、

.....ロシア軍部隊は五日、首都の数力所でドゥダエフ政権部隊と激しい戦闘を繰り返した。

のような事例では部隊と部隊は異なる実体を指し示しているため、項としては区別されなければならない。前者は 8、後者は 5 事例あった。

#### その他 (37 事例)

その他の事例はさまざまな原因が考えられるが、事態性名詞と項が離れており、対立候補の共起スコアのほうが高い事例では、文外の知識を使わなければ解けない問題が 10 事例あった。

米共和党のドール上院議院内総務は四日、ボスニア・ヘルツェゴビナ政府に対する国連武器禁輸決議の一方的解除を求める法案を上院に提出、.....

においては米が「Xガ解除(する)」の項になっているが、これは共和党では決議の解除ができないという知識がなければ米が項であると結論づけることはできない。

残りの 27 事例中、並列構造でいずれを当てても正解としてよいなどすぐ対処できる事例が 15 事例あったが、

引用があり文の構造を考慮しなければならないなど、12 事例は解決が難しい事例であった。

## 5 おわりに

本稿では事態性名詞の項構造解析について項同定に向けた予備実験結果を報告した。事態性名詞の項構造解析に動詞と格要素の共起尺度を用いる手法について比較検討し、共起情報が事態性名詞の項同定に有用であることを示した。また、項同定が共起モデルだけでは決定できない事例を分析し、大部分が構文パターンをうまく捉え切れていないことが分かった。

今後の課題として、今回は共起スコアや構文パターンを素性の一つとして利用したが、どのように解析モデルに組み込むかは未解決な問題である。共起スコアで項同定の名詞候補をフィルタリングする手法や、特定の機能動詞結合の場合はペアによる比較をせず直接決定的に項を決める手法など、いくつかの利用方法が考えられる。どの手順で項構造解析を行うかも含めて検討し、最終的な述語項構造解析器に組み込んだ結果を評価したいと考えている。

## 6 謝辞

Web から取得した 5 億文データを使用させてくださった NICT の河原大輔氏に感謝いたします。

## 参考文献

- [1] Daniel Gildea and Daniel Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, Vol. 28, No. 3, pp. 245–288, 2002.
- [2] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank Project: An Interim Report. In *Proceedings of the NAACL/HLT 2004 Workshop Frontiers in Corpus Annotation*, pp. 24–31, 2004.
- [3] 黒橋禎夫. 京都テキストコーパス Version 4.0., 2005.
- [4] 飯田龍, 小町守, 乾健太郎, 松本裕治. NAIST テキストコーパス: 述語項構造と共参照関係のアノテーション. 情報処理学会研究会報告 (自然言語処理研究会), No. NL-177, pp. 71–78, 2007.
- [5] 小町守, 飯田龍, 乾健太郎, 松本裕治. 共起用例と名詞の出現パターンを用いた動作性名詞の項構造解析. 言語処理学会第 12 回年次大会, pp. 821–824, 2006.
- [6] 藤田篤, 乾健太郎, 松本裕治. 自動生成された言い換え文における不適格な動詞格構造の検出. 情報処理学会論文誌, Vol. 45, No. 4, pp. 1176–1187, 2004.
- [7] Thomas Hoffman. Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR 1999*, pp. 50–57, 1999.
- [8] Daisuke Kawahara and Sadao Kurohashi. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of the LREC 2006*, pp. 1344–1347, 2006.
- [9] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution. In *Proceedings of the COLING/ACL 2006*, pp. 625–632, 2006.
- [10] Vladimir N. Vapnik. *The Statistical Learning Theory*. Springer, 1998.
- [11] 村木新二郎. 日本語動詞の諸相. ひつじ書房, 1990.