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

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

我々は文の包括的な意味解析を行うため，動作性名詞の項構造解析を提言する．動作性名詞とはサ変名詞と動詞由来の名詞であり，これらの名詞は事態を指すとき項構造を持つ．たとえば「彼の決断は正しかった」という文において,「彼が（なにかを）決断（する）」という関係を解析の対象にする．

動作性名詞の項構造解析は，述語項構造解析と同様，文中の項構造を決定し，項を同定する作業の延長と位置づけることができ，情報抽出や質問応答システム，言い換えや機械翻訳などさまざまな分野に応用できる要素技術の一つである．

本論文ではまず動作性名詞の項構造解析とそれに必要なコーパスの作成について紹介し，動作性名詞の項構造解析を事態性判別と項同定の 2 つのタスクに分けた上で，大量のデータを用いて教師なしに文の構造を学習する手法と，動詞と名詞の共起用例を用いて項同定を行う手法とを提案する．

2 動作性名詞の項構造解析

事態性とは文脈中で名詞がコト（動作）を指すかモノ（物体）を指すかという意味的な違いに対応し，動作性名詞の項構造解析とは，名詞に事態性があるとき項構造を決定し，項を同定する解析を指す．文脈に応じて動作性名詞に事態性があるか否か判別する処理を**事態性判別**，項構造を決定して項を同定する処理のことを**項同定**と呼ぶ．項同定という点は従来の述語項構造解析と同様であるが，動作性名詞の場合は，文脈によって名詞が物体を指す場合と動作（**事態**[1]）を指す場合の多義性があるため，動作性名詞が事態であるか否かの判定（事態性判別）を行う必要がある．

3 事態タグつきコーパスの作成

動作性名詞の項構造解析を行うために，我々は京都テキストコーパス [1] を対象に動作性名詞の項構造に関する情報を人手で付与したコーパスを作成中[2] である．

[1]ここで事態性というのは名詞が特定の出来事を指している場合だけではなく，総称的に使う場合も区別せず解析の対象に含める．

[2]述語項構造や名詞句照応，名詞句の関係解析に関する情報も付与している．仕様は http://cl.naist.jp/~ryu-i/coreference_tag.html

このコーパスでは，文章中の各動作性名詞について事態性の有無を判別し，事態性がある場合には項構造（必須格となるガ格・ヲ格・ニ格）の情報を付加している．たとえば

> リスク 管理 の必要性が強く叫ばれているが、市場の 実態 が把握できていないため打つ手がないのが実情。ＢＩＳ が昨年春から 調査 の手法について検討していた。

という記事に対して，

- 管理（する）[ガ:＜外界＞，ヲ:リスク]
- 調査（する）[ガ:実態，ヲ:ＢＩＳ]

のような情報を付与する.「管理」のヲ格「リスク」や「調査」のヲ格「ＢＩＳ」のように，項が文内に出現している場合はそれを形態素単位で指示する．また,「調査」のガ格「実態」のように，文外に出現する項でも記事内で特定できる場合はその要素を指示する．さらに「管理」のガ格のように，必須格で，かつ文内にも記事内にも出現していない場合は＜外界＞タグを付与する．現在のところ，＜外界＞タグについては，一人称（話し手・書き手）・二人称（聞き手・読み手）・それ以外の 3 種類に細分化して付与している．

2006 年 1 月現在，780 記事（6,500 文）に対してタグ付けが完了しており，うち 140 記事は 2 名の作業者によってタグ付与し，タグの一致率を調査し，以下の 4 種類に分類した（表 1）．

	作業者 1	作業者 2	指示先不一致	タグ一致
ガ	3	4	138	604
ヲ	15	54	30	280
ニ	13	9	5	43

表 1: 作業者間のタグの一致率

タグ一致 2 名の作業者間で事態性のタグを付与した動作性名詞とその格要素がいずれも一致した事例

指示先不一致 事態性のタグを付与した動作性名詞と項構造は一致したが，格要素の指示先が不一致だった事例

	文内（同一文節/前文節）	文外 (記事内/外)
ガ	284(18/97)	306(139/167)
ヲ	235(119/69)	46(44/2)
ニ	34(4/13)	6(6/0)

表 2: ガ格・ヲ格・ニ格の分布

作業者 1（作業者 2） 作業者 1（作業者 2）のみが事態性タグを付与した事例

事態性のある動作性名詞は必ずガ格を伴うので，指示先の不一致はあるものの，各動作性名詞に対して事態性の有無はほぼ高い一致率で付与できることが分かる．また，ヲ格とニ格で片方の作業者のみが格要素を指定しているのは，各作業者間で事態の必須格の判断が一致しないということに対応し，ガ格以外の項構造を同定することが比較的難しいことを示している．

一方，2 名の作業者によってタグ付与し，タグの見直しを行った信頼性の高い新聞記事 80 記事を対象に，各表層格（ガ格・ヲ格・ニ格）の分布を調べた（表 2）．

表 2 より，ガ格は文内と文外に広く分布するが，ヲ格は大多数が文内にあり，また比較的動作性名詞の近辺に存在する．つまり，ガ格の項同定を行うには文外の項も正しく当てないと再現率が向上しないが，ヲ格に関しては文内の項を正しく当てるだけでも再現率がかなり高くなる．

4 動作性名詞の項構造解析へのアプローチ

動作性名詞の項構造解析をするに当たり，我々はまず事態性判別を行い，その後事態性のある動作性名詞に限って項構造解析に入るという手順で問題を解く．事態性判別は文中に出現する動作性名詞を事態性あり／なしの 2 クラスに分類する問題なので，事態性に関する曖昧性のない事例を用いて教師なしの学習を行うことができる．そこで，本論文ではこの方法に従って事態性判別のタスクと項同定のタスクを分けて扱う手法を提案し，その有用性を示す．

5 動作性名詞の事態性判別

5.1 手法

事態性のある動作性名詞には,「リスク管理」や「彼の決断」のように，動作性名詞のある文節内や係り受け関係にある文節に項が存在することが多く，こういった名詞の出現パターンを利用することによって事態性判別の精度が向上すると考えられる．そこで，名詞の出現パターンを捉えるための手段として BACT [3] [2] を用いて事態性のある名詞と事態性のない名詞との出現パターンを学習することを考えた．

BACT は文の構造を素性として入れることによって訓練事例の判別に効果が高い構造をルールを学習できる．

[3]http://chasen.org/~taku/software/bact

動作性名詞の分類語彙表 [6] 中での分類項目
「A の B」（例：税の軽減）という表現があったとき,
名詞の分類語彙表中の分類項目の上位 4 桁
動作性名詞が複合名詞であったとき,
各名詞についての分類語彙表中の分類項目の上位 4 桁
動作性名詞の前後 1 文節の形態素列
動作性名詞の出現パターン

表 4: 事態性判別に用いた素性

	精度	再現率
名詞の出現パターンなし	72.3%	58.7%
名詞の出現パターンあり	73.3%	80.2%

表 5: 事態性判別実験結果

5.2 名詞の出現パターンの学習

名詞の出現パターンは，名詞の出現する前後 3 形態素，名詞の出現する文節および係り元の文節の形態素列を木構造にして用いた．事態性に関する曖昧性がない事例として，日本語語彙大系 [3] にサ変動詞として登録されている用言のうち，一般名詞意味属性体系の「名詞-抽象-事-{ 人間活動, 事象 }」ノードの下にあり，かつそれ以外のノードの下にない名詞を正例とした．また，一般名詞として登録されている名詞のうち「名詞-具体」ノードの下にあり，かつそれ以外のノードの下にない名詞と固有名詞を負例とした．学習には新聞記事約 1ヶ月分 [4]（正例:117,581 事例，負例:282,419 事例）を使用した．たとえば「商品取引」の出現パターンは図 1 のような木構造を作成して素性に使い，正例として訓練事例に追加する．

以上のような手順で獲得したルールのうち，重みが高い規則を表 3 で示した．

5.3 実験

事態性の判別には Support Vector Machines [4] [5] を用いて文脈に応じた動作性名詞の事態性を学習し，10 分割交差検定によって事態性の有無の判別性能を評価した．ベースラインには簡単な構造を素性として（「A の B」と複合名詞）を用い，BACT から獲得した名詞の出現パターンを考慮したモデルと比較した．使用した素性を表 4 にまとめた． 評価事例には新聞記事 80 記事（800 文）を用いた．含まれる動作性名詞は 1,237 個（うち 590 個が事態性ありの事例）あった．

表 5 に示すように，事態性のある名詞の出現パターンを用いることによって事態性判別の結果が向上した．名詞の前後 3 形態素と名詞の含まれる文節および係り受けの文節，という比較的局所に限られた形態素列の情報を使うだけでも再現率が大きく上がることが分かったので，ルールとして学習する構造に文節内の依存構造や名詞の意味クラスなどの情報も使うことが考えられる．

[4]http://www.chasen.org/~taku/software/TinySVM

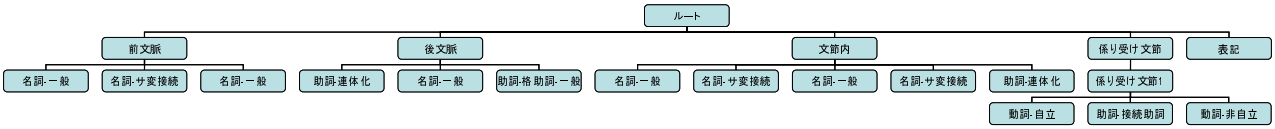

図 1: 「商品取引」の出現パターンの例

事態性ありの判定に効果が高いルール	スコア	事態性なしの判定に効果が高いルール	スコア
同一文節中にサ変名詞+サ変名詞がある	1.13	後ろにサ変名詞がある	- 0.92
後ろに助詞+サ変名詞が続く	1.01	前にサ変名詞がある	-0.54
後ろに名詞の接尾辞がある	0.61	前にサ変名詞，後ろに助詞がある	-0.50

表 3: 事態性判別に有効な素性として獲得したルールの例

5.4 エラー分析

事態性がある事例にも関わらず事態性がないと判別を誤った事例に次のようなものがあった．

項が文外に存在 今年の三が日には、お雑煮を食べたらすぐに、のびのびになっている 受賞 後第一作の執筆に取りかかった。

周辺文脈が一般名詞に近い 「野良黒山の会」のリーダー、木場将弘さん方では、妻の和枝さんらが現地と 電話 のやり取りを続けた。

前者の事例を正しく判別するためには，事態性判別を文外の項の同定も含めて捉える必要がある．また，後者の事例の解析には「電話」と項の共起の情報を活用する必要があると考えられる．

6 動作性名詞の項同定

6.1 手法

動作性名詞に事態性があるときには実際に項がどの名詞句に当たるか同定することが必要だが，本論文ではまず文内に項がある場合に限定し，どの程度項が同定できるか実験を行った．

6.2 実験

動作性名詞のヲ格を同定するために，我々は動詞と名詞の共起用例を機械学習の素性として使用した．動詞と名詞と格助詞の 3 つ組から PLSI によって作成した共起モデル [7] を用い，動作性名詞をサ変動詞とみなしてヲ格候補との相互情報量を計算し，各候補の素性に入れた．その他の素性は飯田ら [8] がゼロ代名詞の先行詞同定モデルで使用しているものを使った（表 6）．

評価事例には動作性名詞のタグ付けを行った新聞記事 80 記事を用い，BACT を使用して共起用例の重み付けも含めた最適な素性を学習し，動作性名詞のヲ格の候補としてもっともふさわしいものをトーナメントモデル [8] によって選択した．BACT には図 2 に示したような構造を素性として与え，共起用例に関する素性は右候補に関する素性と左候補に関する素性各々に加えた．

動詞と名詞のヲ格の共起用例（新聞記事約 20 年分）
日本語語彙大系による動詞と名詞のヲ格の選択制限
動作性名詞とヲ格候補の名詞が何文節離れているか
動作性名詞とヲ格候補の名詞の前後関係
ヲ格候補の文節の機能語
ヲ格候補の格
ヲ格候補の主辞の品詞
ヲ格候補が人間であるか組織であるか

表 6: ヲ格の項同定に用いた素性

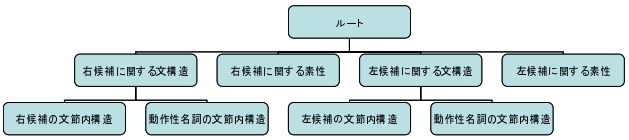

図 2: ヲ格の項同定で使用した BACT の木構造

トーナメントモデルによるヲ格候補の選択では，動作性名詞と実際のヲ格となる名詞，そしてヲ格候補となる名詞の 3 つ組を訓練事例として与える．候補の判定の際には候補となる名詞集合から順次名詞を取り出し，勝ち抜き戦でもっとも勝ち上がった候補をヲ格の名詞であると同定する．各候補の信頼度によって整列させると再現率–精度曲線を描くことができる．

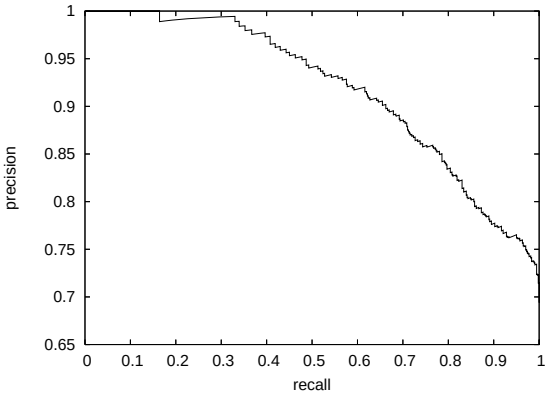

図 3: ヲ格の項同定の再現率–精度曲線

6.3 エラー分析

解析に失敗した事例以下に分析する．

格のルールの影響が強い 医薬品会社「日本商事」株のインサイダー取引 事件$_{正}$ は、大阪地検特捜部が昨年十二月、千葉市の開業医を起訴、日本商事と取引先社員ら計二十四 人$_{誤}$ を略式起訴して 捜査 を終えた。

タグのつけ方の問題 新検査法は、ＨＬＡ$_{誤}$ の 全領域$_{正}$ の 検査 が可能。

ヲ格の共起用例が有効に効かない 数年前に 新築 なった 市$_{誤}$ の 図書館$_{正}$ は、……

1 番目の問題は，ガ格の先行詞同定モデルをそのまま動作性名詞の項構造解析に適用するだけではなく，動作性名詞の項構造解析に向けて調整する必要がある．

また，2 番目の問題は，表 1 にも示したように作業者間でも指示先の不一致が少なからずあるため，名詞の関係を含めたタグ仕様を再検討中である．

最後の問題は，共起用例の素性をノードに含むルールが誤った候補に有利に働いて間違えているので，トーナメント候補のペアでどちらのほうが動作性名詞との共起が強いかといった素性として共起用例を使うなど，トーナメントモデルと共起用例の組み合わせ方を検討するつもりである．

7 関連研究

動作性名詞の項構造解析用の資源としては他に NomBank [9] がある．NomBank では，Penn Treebank [10] に対し，英語における動詞の名詞化に着目して，動詞についての述語項構造解析用のコーパス PropBank [11] の仕様に従って項構造を付与しており，この点では我々のコーパスと似通っているが，NomBank は項の対象が文内に現れる場合にのみ項情報を付加しているので，我々のコーパスのように文外に項候補が現れる場合を考慮していない．事態性判別には項が文外の要素を指しているという情報が必要なので，項が文外にあったとしても情報を付与する必要がある．また，NomBank を用いた解析としては名詞句の項構造には一定のパターンがあるという知見をまとめた Meyers ら [12] の研究がある．

一方日本語に関する名詞句の関係解析としては笹野ら [13] ・河原ら [14] が名詞の格フレーム辞書の構築とコーパスの作成を行っているが，彼らは名詞の項構造を広く捉え，位置や属性，親族関係なども解析の対象としているが，我々は動作性名詞に焦点を当てて解析し，動作性名詞に特化した出現パターンを用いている点が異なる．

8 おわりに

本論文で，動作性名詞の項構造解析のためのコーパス作成と，動作性名詞の項構造解析について述べた．本論文では項構造解析のタスクを 2 つに分け，名詞の出現パターンを用いた動作性名詞の事態性判別手法と共起用例を用いた動作性名詞の項同定手法を提案し，有用性を示した．

事態性判別は精度 76.6%，再現率 79.6%で行うことができ，ヲ格は 71.5%の精度で項を同定することができるので，文内の範囲ではある程度実用的に項構造解析を行うことができると考えられる．今後は文外の候補を同定するモデルを作成し，ヲ格以外の格（ガ格・ニ格）についても項の同定を行い，全体としての評価を行うことを予定している．

また，現在は表層格を対象にした項構造解析を行っているが，語彙概念構造（LCS）に従って編集された項構造の辞書作成が進行中 [15] であり，将来的には LCS を用いた項構造解析に取り組みたい．

参考文献

[1] 黒橋禎夫. 京都テキストコーパス Version 4.0.

[2] Taku Kudo and Yuji Matsumoto. Boosting Algorithm for Classification of Semi-Structured Text. EMNLP, 2004.

[3] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳文, 林良彦. 日本語語彙大系. 岩波書店, 1997.

[4] 毎日新聞. 毎日新聞社, 2002.

[5] Vladimir N. Vapnik. The Statistical Learning Theory. Springer, 1998.

[6] 国立国語研究所. 分類語彙表. 大日本図書株式会社, 2004.

[7] 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, 2004

[8] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌. Vol. 45, No.3, 2004.

[9] Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young and Ralph Grishman. The NomBank Project: An Interim Report. In Proc of the Workshop on Frontiers in Corpus Annotation, HLT-NAACL, 2004.

[10] Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, Vol.19, 1993.

[11] Martha Palmer, Dan Gildea and Paul Kingsbury. The Proposition Bank: A Corpus Annotated with Semantic Roles. Computational Lingustics, 31:1, 2005.

[12] Adam Meyers, Ruth Reeves and Catherine Macleod. NP-External Arguments: A Study of Argument Sharing in English. In Proc of the Workshop on MWE, ACL, 2004.

[13] 笹野遼平, 河原大輔, 黒橋禎夫. 名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析. 自然言語処理, Vol.12, No.3, 2005.

[14] Daisuke Kawahara, Ryohei Sasano and Sadao Kurohashi. Toward Text Understanding: Integrating Relevance-tagged Corpus and Automatically Constructed Case Frames, In Proc of the 4th LREC, 2004.

[15] 竹内孔一, 乾健太郎, 藤田篤, 竹内奈央, 阿部修也. 分類の根拠を示した動詞語彙概念構造辞書の構築. 情報処理学会研究会報告（自然言語処理研究会）, Vol.2005, No.94, pp.123-130, 2005.