述語項構造を意識した名詞データの構築

竹内 孔一 (岡山大学大学院自然科学研究科)[1]

宮田 周 (岡山大学工学部)

河村 一希 (岡山大学工学部)

Construction of Japanese Noun Data on the Basis of Predicate-Argument Thesaurus

Koichi Takeuchi (Graduate School of Natural Science and Technology, Okayama University)
Syu Miyata (Faculty of Engineering, Okayama University)
Kazuki Kawamura (Faculty of Engineering, Okayama University)

要旨

本発表者は日本語の述語項構造辞書を構築し，公開してきた．そこでは，共通概念を約 1200 程度に定義し，意味役割を 31 種類，細分類で 72 種類定義した．これらをもとに，名詞に関する述語項構造辞書構築のための基本データを 2 種類構築している．1 つは非飽和名詞に関する辞書で最終的には，影山 (2011) が提示する Generative Lexicon の構造を予定している．現段階では，非飽和名詞に対して例文を 2500 文作成し，その全てに対して意味役割を付与した．この作業における問題点や作成された例の質について説明する．さらに「相違がある」と「異なる」が同義であるように，述語と言い換えができる名詞表現がある．これらの類語を類語辞典を参考に人手により作例を構築して作成している．人手による作業の結果,「暇を出す」など慣用句表現に近いものが多く獲得できたことを報告する．

1　はじめに

本研究グループでは日本語の述語項構造に対してシソーラス形式で語義毎に例文を作成し，意味役割と語義概念を付与した事例を構築し公開している[2]．この辞書を拡張する形で，名詞の項構造に関する 2 種類のデータを構築しているので報告する．

ひとつは，言語学において分析されている名詞の項構造 (西山 (2003, 2013); 影山 (2011); 庵 (2007); Pustejovsky (1995); Meyers et al. (2004)) である．名詞の項構造は「その芝居の主役」や「彼の上司」における「主役」や「上司」のように密接に関連する語 (ここでは「芝居」,「彼」であり項と考える) を必要とする語である．言語処理の観点からすると NTCIR の RITE-2 含意認識タスクにおいて例えば

(t1) BLT サンドイッチとは、サンドイッチの一種であり、パンに挿む食材として、ベーコン、レタス、トマト が用いられることから、それぞれの頭文字を取って名づけられた。

(t2) サンドイッチの略称として具材となるベーコン、レタス、トマトの頭文字 BLT が用いられるものがある。

の場合,「一種」「略称」「頭文字」といった言葉が項を要求し，これらの関係を解くことが含意認識を解くことに結びつく (竹内 (2014))．

もう一つのデータは名詞まわりの連語である．例えば「考案する」に対して「着想を得る」などの異品詞間での言い換えデータである．これらデータをどのように構築し，現段階でどの程度集まり，どのような問題があるか次章以降で記述する．

[1] koichi@cl.cs.okayama-u.ac.jp
[2] 述語項構造シソーラス (http://pth.cl.cs.okayama-u.ac.jp).

2　名詞の項構造データの構築

2.1　作成するデータの構造

最初の段階として文献 (竹内 (2014)) に記述したように，名詞と名詞が取る例文を作成し，述語項構造シソーラスの意味役割を付与する．例文のタイプとして現段階では「X の Y は Z」の構文をベースとする．Y が対象とする名詞であり，例えば「創立者」では

[あの図書館]【主体】の創立者は [田中さん]【対象（人)】だ

のようになる.「創立者」の項として「あの図書館」と「田中さん」があり，その意味的関係を表すラベルとして【】内に意味役割を付与する[3]．こうした例文ベースの名詞項構造のデータ構築は英語では NomLex(Meyers et al. (2004)) で行われている．一方で，先行研究として日本語における名詞格フレーム辞書 (笹野他 (2005)) では対象名詞と項の事例の大規模収集に焦点がおかれているため例文は存在しない．しかし名詞の項構造に対して例文ベースで行うことには 2 つの利点があると考えられる．一つ目の利点は項構造データ構築の際に人間が正しく関係を記述しやすいと考えられる点である．これはデータ構築の際に単語のペアを付与する場合[4]と，文として成立する表現を一度考えてから項を同定するのでは，あきらかに，後者の方が人間の言語直感を引き出せると考えられる．二つ目の利点は，名詞項構造の自動付与を視野にいれると例文は機械学習における事例として都合が良いことである．

次にこうした例文ベースのデータから最終的な名詞の項構造を表す Generative Lexicon ベースへの構造 (影山 (2011)) との比較を行っておく.「創立者」の場合には下記の様になる．

	「創立者」
外的分類	人間 (x)
目的・機能	
成り立ち	機関 [w] を創立する ‖ 創立 (x,w)

ここで機関 [w] が先ほどの【主体】にあたるもので,「創立者」 は結局，人間のことを表す部分が例文での【対象（人)】である．また「成り立ち」 の項目では動詞「創立」の項としてこれらの要素が結び付けられる.「創立」は既に述語項構造シソーラスに登録されており，概念と意味役割，さらに例文が定義されている[5]．こうした最終構造と例文を比較すると，例文から対象となる名詞のカテゴリ (先ほどの例では「人間」や「成り立ち」) での項の具現化部分が取り出せる．自動で最終構造は作成できないが，半自動で最終構造が得られる見通しである．

2.2　名詞項構造データの構築作業

上記で説明した例文ベースの事例データを構築するには，1) 対象とする名詞のリストの構築，2) 名詞に対する例文の構築，3) 例文に対する意味役割の付与を行う必要がある．以下，順に説明する．

対象とする名詞リスト

付与対象の名詞は項を持つ名詞であるが，どの名詞が項を持つかというのは前もってわからない．よってまず西山 (2003, 2013) に記載されている非飽和名詞，譲渡不可能名詞をリスト化して登録する．次に，NTCIR の RITE1 と RITE2(含意認識タスク) の開発データ例文すべてを形態素解析して，名詞に該当するものをすべて登録する．これは作成した名詞項構造データの評価として含意認識タスクを利用することを想定しているためである．優先順位としては文献から獲得した名詞リストを先にすることで，確実な非飽和名詞・譲渡不可能名詞のデータを構築する．RITE-2 から得られた名詞のリストには項構造を持たない対象外の名詞も含まれる．よって作業者は不要な名詞を分ける作業を行

[3]意味役割の全体系について簡単な説明が竹内 (2014) にある．
[4]ここで単語のペアの付与とは例えば直接項構造を作業者に記述させるような付与タスクである．
[5]Web サイトで検索して確認できる (http://pth.cl.cs.okayama-u.ac.jp)．

う必要が出てくる．

例文の構築

上記で決定した付与対象候補の名詞のリストに対して「X の Y は Z だ」の例文を作成する．各名詞に対して例文を作成し，後の意味役割付与などのデータ管理を行うためにブラウザベースの作業システムを CakePHP を利用して作成した．作業結果は MySQL に保存できるため，MySQL データを確認することで進捗を確認することが容易になる．

例文の作成において,「X の Y は Z だ」の構文には制約があり，Z は必ず名詞になるように表現する．例えば,「その演劇の主役は太郎だ」のように「太郎」など具体的に入れることで,「主役」は人間であることなどがわかる．これが Z に形容動詞などを許すと「その演劇の主役は立派だ」など表層的には適合しているが，必要とする情報が得られないためである．

しかしながら一方で，項構造がある名詞であるがこの構文では Z を具体的に表現できない場合がある．例えば譲渡不可能名詞「鼻」では「象の鼻はそれだ」となる．これは Z が具体例の名前を求めているためであり，無名のインスタンスでは表現することができず,「それ」などの指示詞でしか表現できない．非飽和名詞でも同様で例えば「理由」では「あの行動の理由はそれだ」という表現になる．現状ではこうしたインスタンスの名前が無い場合の名詞に対してどのような構文を適応すればよいか自明でないため，現段階では「それだ」ではなく例えば「美しい」など作業者が自然だと思う例文を構築している．

意味役割の付与

作成された例文に対して意味役割を付与する．CakePHP による作業システムは例文が作成されると，MeCab による形態素解析を行い，形態素単位に分割して，意味役割の付与が行えるようにする．意味役割の体系は述語項構造シソーラスに準拠するがほとんどの場合,【主体】と【対象】の付与となる．

2.3　名詞項構造データの付与作業結果と考察

対象とする名詞のリストであるが，文献から得られた名詞は 66 語，含意認識タスクから自動で獲得した名詞は 16774 語である．次に例文の付与であるが，学部学生 2 名の作業者に例文を付与していただいた．その結果 2532 事例登録できた．作業から例えば「出身」(「太郎の出身は岡山県だ」) など新たな名詞の項構造例文が付与できている．

一方で，全てが正しい例文ではない．例文を作成する段階で作業ミスがいくつか見受けられる．例えば「花」の例文で「その花はきれいだ」など「花」にかかる項の部分を全く記述せずに表層的に「X の Y は Z だ」に当てはめてしまっている．これは作業者が言語データ付与に未経験であること，また分野としても言語とは関係無かったことが原因として考えられる．また，今回の作業枠組では対応できていないことも原因である．この例ではまず「花」の語義から分類して (植物の花または職場の花など)，次に項として必須となるもの (「植物」や「職場」の具体例) を検討する必要がある．

次に意味役割付与についてであるが 3199 箇所 (約 2500 例文) 付与できている．意味役割の付与作業は例文を作成した作業者とは別で，BCCWJ の意味役割付与を行った作業者が付与した．付与した意味役割のラベルの揺れを確認するために部分的にではあるが別の付与作業者 (BCCWJ の意味役割付与を行った作業者) に付与をお願いしており，現在その結果を分析中である．基本的には意味役割の細分類，つまり【対象（人)】か【対象（生成物)】かなどどういう分類でアノテーションされているかが名詞項構造データを構築する上で重要となる．このあたりを中心に分析をすすめたい．

これに関連して，名詞の項構造の例文と意味役割付与を行うなかで問題となっているのが，名詞の概念カテゴリの必要性である．例えば,「主役」の場合には,「その演劇の主役」のように「X の Y」における X は「演劇などの名詞」がくる．こうした選択制限情報はのちの言語処理では有効と考えられるが必要とされる名詞概念の粒度の予測が立っておらず付与できていない状態である．当然，例文中に「その演劇」とインスタンスで記しているので，これらをもとに類似度計算などでの処理は可能である．

さらに名詞の基本情報として語義が必要である．京都大学名詞格フレーム辞書には国語辞典と規

則から作成した語義に相当するラベルが格スロットとして付与されている．例えば「ドリル」なら「工具」か「演習，問題」かである．ただ自動獲得であるため誤りも少なからず存在し，語義を辞書ベースで分けて付与すべきか，自動獲得ベースのデータを整理して付与すべきか方針がまだ固まっていないのが現状である．

3　名詞まわりの連語

名詞まわりの連語を獲得するために，類語辞典から述語の類語を探し，人手で例文を付与することで連語のデータを構築する．類語辞典としては角川類語辞典を選び，述語項構造シソーラスの述語と類語辞典との単語のマッチングを行い，対応する類語の分類から述語に対する類語候補を獲得した．これをもとに人手で言い換えとなっている語を抽出し，連語表現を作成した．下記の表に獲得した例を示す．

連語	シソーラスの述語	例文
違いがある	異なる	報道と事実に相違がある
着想を得る	思いつく	漫才師がネタの着想を得る
手抜かりがある	荒っぽい	仕事に手抜かりがある
焼き餅を焼く	妬ける	周囲が二人に焼き餅を焼く

アノテーション作業により現在 100 語ほど獲得できている．各例文には意味役割付与を行っている．

4　まとめ

述語項構造シソーラスの体系を利用して，名詞に関連した項構造データと連語データの構築を行っている．意味役割ラベルと語義概念を一貫して構築できるのが利点である．現段階では項構造では約 2500 の例文を構築して，意味役割付与が一人の作業者で付与できた段階である．今後，項構造のデータの評価ならびに拡張，連語データの拡張を行う予定である．

謝辞

本研究は，科研費 (26370485) の助成を受けたものである．

文献

Adam Meyers, Ruth Reeves, and Catherine Macleod (2004) “NP-External Arguments: A Study of Argument Sharing in English,” in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pp. 96–103.

James Pustejovsky (1995) *The Generative Lexicon*: MIT Press.

庵功雄 (2007) 日本語におけるテキストの結束性の研究，くろしお出版．

影山太郎 (2011) 日英対照 名詞の意味と構文，大修館書店．

笹野遼平，河原大輔，黒橋禎夫 (2005)「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，第 12 巻，第 3 号，pp.129–144．

西山佑司 (2003) 日本語名詞句の意味論と語用論，ひつじ書房．

西山佑司（編） (2013) 名詞句の世界，ひつじ書房．

竹内孔一 (2014)「述語項構造シソーラスを意識した名詞の意味構造アノテーションのための名詞意味構造の検討」，第 6 回コーパスワークショップ予稿集，pp.51–56．