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

笹野 遼平　　河原 大輔　　黒橋 禎夫
東京大学工学部　東京大学大学院情報理工学系研究科
{sasano, kawahara, kuro}@kc.t.u-tokyo.ac.jp

1　はじめに

文章中の離れている要素の関係付けを行うことは自動要約、機械翻訳、質問応答などの言語処理アプリケーションを高度化するために必要となる。このような関係付けの一つに名詞句の関係付けがある。例えば「チケットを買ったら、値段は 2000 円だった。」という文では「チケット」と「値段」を関係付けることが必要となる。本論文では、その名詞にとって必須的な関係をその名詞の必須格と呼ぶことにする。

名詞の必須格の多くは「チケットの値段」などの形で出現する。しかし、文脈的に明らかな場合は必須格は明示されないことがあり、その場合は必須格にあたる語が何であるかを解析する必要がある。その際、その名詞がどのような必須格をとるのか、各必須格にはどのような用例があるのかという知識が必要となる。それらを記述した名詞格フレーム辞書は今のところ存在せず、また、人手で作成することはほとんど不可能であるためコーパスから自動構築することを考える。

動詞・形容詞など用言についての格フレーム辞書はコーパス中に表層的に明示された格情報とその用例を用いて自動構築する手法がすでに提案されている [1]。名詞で同様の手法を用いる場合、ノ格として対象の名詞に係る用例を集めることとなる [4]。ただし、用言の場合はガ格やヲ格といった格の種類によって対象の語との関係がある程度制限されるのに対し、名詞の場合、ノ格により 2 つの名詞が関係を持つことが明示されていてもそれらの持つ関係は様々であるため、名詞格フレーム辞書を構築する場合には名詞句の関係の解析を行うことが必要となる。

本論文では、コーパスから収集した名詞句「AのB」の意味解析を国語辞典の定義文を利用して行い [3]、その結果を用いて名詞格フレーム辞書を自動構築する手法を提案する。また、自動構築した名詞格フレーム辞書を用いてテキスト中に出現する名詞句の関係解析を行う手法を提案する。

2　名詞句「AのB」の意味解析

2.1　国語辞典を用いた意味解析

国語辞典の定義文には、その語に必須的な要素が含まれていることが多い。国語辞典に出現するこのような要素に注目することによって、日本語の自動解析で困難な問題とされている名詞句「AのB」の意味的曖昧性を自然に解決することができる。例えば、『例解小学国語辞典』[6] における「値段」の定義文は次のようになっており、「品物」という必須要素が含まれている。

【値段】品物 を売り買いするときの金額。

このような定義文に含まれる必須要素との対応付けを行うことにより、「チケットの値段」という名詞句は「チケットを売り買いするときの金額」という意味であると解析できる。また、多義性を持つ語であれば複数ある定義文のうち対応する定義文との関連付けを行うことにより多義性を解消することも可能となる。

2.2　意味属性を用いた意味解析

国語辞典中に適当な定義文のない語や、パターン化した解析が適した句があるため、国語辞典による解析の他に、シソーラス [5] の意味属性に関して設定したルールを用いた意味解析を行う。設定したルールに基づいた解析によって、例えば「彼の母」には<必須格（親族）>、「赤色の帽子」には<修飾>の関係が与えられる。以下にルールの例を示す。

- A:《人》, B:《親族関係》 → <必須格 (親族)>
- A:《＊》, B:《場》 → <必須格 (位置)>
- A:《数量》or《色》, B:《＊》 → <修飾>
- A:《時間》, B:《＊》 → <時間>
- A:《組織》, B:《主体》 → <所属>
- A:《主体》, B:《＊》 → <所有>

3　名詞格フレーム辞書の自動構築

3.1　名詞句「AのB」の収集

コーパスから普通名詞Bにノ格で係る名詞Aの用例を収集する。表層的に「AのB」の形をしているもののうち「AのBが」や「AのBを」のようにBがガ格やヲ格であるものやBが文末に来ているものなどAがBに係っている可能性が高いものを集める。

3.2　用例の解析

収集された「AのB」に含まれる各Bについて集まったAの用例を解析し、解析結果ごとにまとめる。以下ではそれぞれのまとまりを格スロットと呼ぶ。解析結果のまとめ方は次のとおりである。

- 定義文中の語に関係付けられた用例は定義文中の語ごとにまとめる。ただし、定義文中で並列に出てくる語に集まった用例はひとつにまとめる。
- 意味属性による解析で関係が得られたものはその関係でまとめる。
- 意味解析結果の得られなかったものは<その他>としてまとめる。

3.3　格スロットの選択

用例の解析結果をまとめた格スロット全てをそのまま格フレーム辞書に含めてしまうと、必須的でない格スロットまでもが格フレーム辞書に含まれてしまう。このため、格フレーム辞書に含める格スロットを絞りこむ必要がある。

必須的な格スロットであれば、その用例も多いと考えられる。また、定義文中に含まれる語に関連付けれられた格スロットや、意味属性を用いた解析で<必須格>と解析された格スロットは必須要素である可能性が高いと考えられるなど、格スロットの種類によって必須要素である可能性が異なると考えられる。したがって、格スロットの種類ごとに必須格とみなす閾値を設定する。設定した閾値を表1に示す。

表 1: 必須格スロットと判断する閾値

格スロットの種類	用例の出現頻度
定義文に関連付け	0.5 % (1/200)
<必須格>	2.5 % (1/40)
<所属>	2.5 % (1/40)
<所有>・<主体>・<場所>	5 % (1/20)
<その他>	10 % (1/10)
<修飾>・<時間>	使用しない

出現頻度:Bの出現回数に占める「AのB」の割合

表 2: 名詞格フレームの例

見出し	格スロット	用例
表情 (1)	〔自分〕 〔気持ち〕	<人>人々, 相手,… <動作>安ど, 余裕,…
レバー (1)	〔動物〕	牛, 鶏, ブタ
レバー (2)	〔機械〕	トランク, 機械, 装置,…
引き出し (1)	〔机・たんす〕	机, タンス, 鏡台,…
引き出し (2)	<その他>	預金, 資金, 貯金,…
コーチ (1)	〔スポーツ〕 <所属>	野球, 水泳, 体操、… <組織>チーム, 部,…
株式 (1)	〔会社〕	<組織>企業, 会社,…

【表情】自分 の 気持ち を顔や身ぶりにあらわすこと。また、その顔つき。
【レバー】① 動物 のかんぞう。
② 機械 を運転するために、手でにぎる棒。
【引き出し】机・たんす などにある、引いて出し入れができる箱。
【コーチ】スポーツ で、そのやり方などを教えること。また、その人。
【株式】株式 会社 が仕事を始めるとき、元手となるお金を多くの人に出してもらうために分けた一つ一つ。

3.4　格フレームの構築

格フレームとは、ある語のとり得る格の制約を記述したものであり語義ごとに必要となる。基本的に定義文ごとに1つの格フレームを構築する。ただし、<必須格>や<その他>としてまとめられた格スロットについては、対応する適切な定義文が存在しなかったために作成された格スロットであると考え、新たな格フレームを構築する。格フレームの構築方法を以下にまとめる。

1. 表2の「表情」のように、同一の定義文に関連付けられた格スロットが複数ある場合は同一の格フレームの別の格スロットとして扱う。
2. 表2の「レバー」のように、別の定義文に関連付けられた格スロットがある場合はそれぞれ別の格フレームを構築する。
3. 意味属性を用いた解析で<必須格>と分類されたか、<その他>としてまとめられた格スロットがある場合は新しい格フレームを作成する。
   例えば表2の「引き出し」の場合、定義文に関連付けられた格スロットの他に、<その他>を格スロットとして持つ新しい格フレームを作成する。
4. 意味属性を用いた解析結果で<必須格>以外に分類された格スロットがある場合は既に存在している格フレームに付け加える (他の必須格スロットがない場合は新しい格フレームを作成する)。
   例えば表2の「コーチ」の場合、<所属>と分類された格スロットは定義文に関連付けられた格フレームに付け加える。

4　名詞句の関係解析

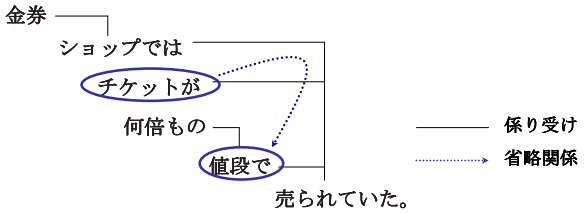
自動構築した名詞格フレーム辞書を用いた名詞句の関係解析の方法について説明する。名詞の関係解析とは対象の名詞 (照応詞) が照応している先行詞を特定する処理である。解析は入力文の前から順に用言の省略解析と並行して行う。また、そこまでの解析の結果として出力された語も候補語として用いる。名詞の格解析、省略解析の手順は次のとおりである。ただし、複数の格フレームが存在する場合はそれぞれの格フレームについて解析を行い、対応付けの類似度のもっとも大きい格フレームに決定する。

1. JUMAN、KNP を用いて入力文を形態素・構文解析する。
2. 解析対象となる名詞Xに直接係る名詞Aがあるときは、格スロットの用例との類似度を計算する。閾値α (現在は 0.6) 以上となる用例があれば、もっとも類似している用例を含む格スロットに対応付ける。
3. 空いている格スロットがある場合は先行詞を探す。
   (a) 名詞Xと同じ表記の名詞が 2 文前までに出現しており、その関係解析結果が存在する場合は解析結果を引き継ぐ（共参照)。
   (b) 空き格スロットに対して、自分の節、主節、主節の子孫、前文、2 文前の順でそれぞれに含まれる名詞との類似度を計算し、類似度が閾値β (現在は 1.0) 以上であれば格スロットに対応付ける。

例として、図 1 のような文を考える。まず、「金券」に対する格フレームは構築されていないので、「金券」の必須要素はないと判断される。「ショップ」については、直接係っている「金券」が閾値αを満足するので格スロット〔商品〕に「金券」が対応付けられる。

次に、「チケット」の解析が行われる。直接係っている名詞がないことから、先行詞の候補として「値段」、「ショップ」、「金券」の順に格スロットに含まれる用例との類似度が計算される。しかし、閾値βを満足するものがないため必須要素はないと判断される。

最後に、「値段」の解析が行われる。「値段」には「何倍」が直接かかっているが閾値αを満足しないので他の要素から先行詞を探す。この場合、まず「チケット」と格スロット〔品物〕の用例との類似度が計算され。用例中に「チケット」があることから閾値βを満足するので格スロット〔品物〕にチケットが対応付けられる。



見出し	格スロット	用例	解析結果
ショップ	〔商品〕	商品, 雑貨,…	金券
チケット	〔乗り物・劇場〕	公演, 試合,…	なし
値段	〔品物〕	品, チケット ,…	チケット

【ショップ】商品 をならべて売る所。
【チケット】乗り物 や 劇場 などで、料金をはらったしるしにくれる紙のふだ。
【値段】品物 を売り買いするときの金額。

図 1: 関係解析例

5　実験と考察

5.1　自動構築された辞書の評価

毎日新聞 12 年分および日経新聞 13 年分の約 2,500 万文を用いて名詞格フレーム辞書の自動構築を行った。約 17,000 語の名詞について格フレームが構築され、1 語あたりの格フレーム数の平均は 1.06 個、1 つの格フレームに含まれる格スロット数の平均は 1.09 個であった。

コーパス中に 1 万回以上出現した普通名詞から無作為に抽出した 100 個の名詞について正しいと考えられる格フレームを人手で与え、それらと自動構築された格フレーム辞書を比較することにより自動構築された辞書の評価を行った。その結果を表 3 に示す。必須格スロットが完全に一致しているものを正しい格フレームとして評価している。

5.2　関係解析実験

新聞 10 記事を用いて名詞の関係解析の実験を行い、解析システムの出力結果と人手で付けられた正解 [2] を比較することにより解析システムの評価を行った。その結果を表 4 に示す。適合率、再現率は必須要素が照応詞と直接係り受けが無い場合の解析精度を評価したものである。主な誤り原因を以下に示す。

表 3: 格フレームの精度

適合率	再現率	F
58/70 (0.829)	58/68 (0.853)	0.841

表 4: 関係解析の精度

適合率	再現率	F
31/60 (0.517)	31/46 (0.674)	0.585

必須要素の文脈依存性

解析システムの誤った出力には、文脈によって格スロットの必須性が異なることに起因するものがある。

(1) 子会社の 株式 を売却した。

この例の場合、「株式」とは「子会社の株式」であり、どこの会社のものかという情報が必須となる。このような用例があるため「株式」の格フレームが構築されている (表 2 参照)。

(2) 株式 相場の押し上げ要因となる。

ところが、この例の場合、「株式」にはどこの会社のものであるかという情報は必要でない。このような違いを判断できないため、この例のような場合でも〔会社〕にあたるものを探してしまう。

適切な用例の不足

必要な先行詞が出力されないものには、必要な格スロットが構築されないことに起因するものがある。

(3) 韓国の金大中 政権 への期待が高まる。

例えば「政権」の場合、<国>にあたるものが必須的要素であると考えられる。ところが、この例のように「<国>の<人>政権」という形で通常用いられるため、「<国>の政権」の用例は少なく、<国>に対応する格スロットは構築されない。

同じ意味属性をもつ名詞の存在

必須要素の取り違えには、より近くに同じ意味属性を持つ名詞が存在することに起因するものがある。

(4) ブッシュ**米**政権は外交・安保政策で「単独主義」を意識的に控え、国際社会との協調路線を打ち出した。たが、**ロシア**との弾道弾迎撃ミサイル制限条約からの一方的脱退宣言や、地下核実験の将来の再開を示唆する姿勢にはブッシュ大統領 らが繰り返し主張している国益優先の哲学が色濃く反映されている。

このような文章があると、2 文目に出現する「大統領」という名詞には「<国>」という要素が必須であることが自動構築された格フレーム辞書からわかるので、システムは「<国>」にあたるものを探す。ところが「米の大統領」と解析すべきであるのに、もっとも近くにある<国>は「ロシア」であるため、「ロシアの大統領」であると解析してしまう。この問題を解決する方法としては、1 文目に出現する「ブッシュ米政権」というトピック的な情報を用いることや、社会・文化に対する知識をシステムに持たせることなどが考えられる。

6 おわりに

本論文では、名詞句「AのB」に注目することによりコーパスから名詞格フレーム辞書を自動構築する手法、および自動構築した名詞格フレーム辞書を用いて名詞句の関係解析を行う手法を提案した。

名詞格フレーム辞書を自動構築した結果、約 17,000 個の名詞について格フレームが構築された。適合率は 82.9%、再現率は 85.3%であり、ある程度実用的な名詞格フレーム辞書を構築できたと考えられる。また、自動構築した名詞格フレーム辞書を用いて関係解析を行ったところ、適合率 51.7%、再現率 67.4%の精度で解析できた。

今後、名詞格フレーム辞書に含める必須要素を集める際に複合名詞や「AというB」、「Aに関するB」などの用例も集めることによりカバレージのさらに大きい名詞格フレーム辞書を作成し、また、名詞句の関係解析方法についても改良を加えていく予定である。

参考文献

[1] Daisuke Kawahara and Sadao Kurohashi. Fertilization of Case Frame Dictionary for Robust Japanese Case Analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 425–431, 2002.

[2] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese Relevance-tagged Corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013, 2002.

[3] Sadao Kurohashi and Yasuyuki Sakai. Semantic Analysis of Japanese Noun Phrases : A New Approach to Dictionary-Based Understanding. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, pp. 481–488, 1999.

[4] Masaki Murata, Hitoshi Isahara, and Makoto Nagao. Resolution of Indirect Anaphora in Japanese Sentences Using Examples “X no Y”(Y of X). In *Proceedings of ACL’99 Workshop on ‘Coreference and Its Applications’*, 1999.

[5] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[6] 田辺洵一（編）. 例解小学国語辞典. 三省堂, 1997.