E-009

名詞と助数詞の呼応関係に基づく助数詞オントロジーの自動構築

Automatic Construction of Ontology of Japanese Classifier based on Agreement between Nouns and Classifiers

白井 清昭 *
SHIRAI, Kiyoaki

徳永 健伸 †
TOKUNAGA, Takenobu

1 はじめに

日本語では名詞を数える際には一般に助数詞を必要とし，助数詞の種類も豊富である．さらに，例えば犬は「匹」では数えられるが「個」では数えられないように，ある名詞を数える際には特定の助数詞のみが使われるといった名詞と助数詞の呼応関係が存在する．計算機で助数詞を適切に取り扱うためには助数詞に関する知識の整備が必要不可欠である．

本研究は，日本語の助数詞オントロジーを自動的に構築することを目的とする．助数詞オントロジーとは，助数詞をその一般性に応じて体系的に整理した知識ベースである．例えば，「頭」は比較的大型の動物を数えるときに使われるのに対し，「匹」は動物全般を数えるときに使われる．したがって，「匹」は「頭」よりもより一般的な助数詞であるといえる．このような助数詞の一般性の違いを通常のオントロジーと同様に木構造で表現した知識体系を構築することが本研究の目標である．助数詞オントロジーは，自然言語解析や生成のための有用かつ基礎的な言語資源になりうる．また，日本語教育の面からも，外国人にとって日本語の助数詞の使い方を覚えることは難しいとされているため，日本語教育教材としての活用も期待できる．

2 関連研究

名詞と助数詞の呼応関係に関する研究のひとつに飯田の研究がある [3]．飯田は，33個の主要な助数詞の意味と用法をインフォーマント調査や日本語テキストの調査などを通じて分析し，名詞を数える際に用いる助数詞の選定プロセスを明らかにした．Bondらは，機械翻訳における文生成に利用するという前提で，シソーラスにおける名詞の意味クラスを利用し，個々の名詞に対して生成するべき適切な助数詞を効率良く選択する手法を提案した [1]．他に，助数詞の用法を比較言語学の観点から分析した研究がいくつかある [5, 6] が，日本語の助数詞に関する研究はそれほど盛んに行われてきたわけではない．少なくとも本研究のように助数詞の一般性に着目して助数詞オントロジーを構築する試みは存在しない．

本研究と同様に，名詞と助数詞の呼応関係に基づいてオントロジーを構築する試みとしてHuangらによる研究がある [2]．Huangらは中国語の名詞のオントロジーを構築したのに対し，本研究では助数詞のオントロジーの構築を目指す点が異なる．

3 提案手法

3.1 上位-下位関係の獲得

本項では上位-下位関係にある助数詞の組を自動的に獲得する方法について述べる．ここでは助数詞の上位-下位関係を以下のように定義する．助数詞 c_1 と呼応する名詞が c_2 と呼応する名詞よりも一般的あるいは同等の概念を表わすとき，c_1 は c_2 の上位の助数詞であるとする．また，助数詞の上位-下位関係を $c_1 \succ c_2$ と表わす．例えば，「店」は店を数える助数詞，「軒」は店を含む建物一般を数える助数詞なので，「軒」は「店」の上位の助数詞である．

本研究では，名詞と助数詞の呼応関係を基に助数詞の上位-下位関係を自動的に獲得する．まず，呼応関係にある名詞 n と助数詞 c の組 (n, c) を集めたデータベースを構築する．このデータベースは，様々な名詞とそれらを数える際に用いられる助数詞を記載した辞典 [4] から (n, c) を書き起こして作成した．呼応関係にある名詞と助数詞の組の数は9,582組であった．さらに，1種類の名詞としか呼応しない230個の助数詞については，これらは特殊な用法であるとみなして除去した．最終的に9,352組の (n, c) を含むデータベースが得られた．データベースに含まれる名詞と助数詞の異なり数はそれぞれ4,624，331であった．

次に，助数詞と呼応関係にある名詞の集合の包含関係に着目する．いま，助数詞 c_k と呼応する名詞の集合を N_k とおく．もし，2つの助数詞 c_i と c_j について，N_i が N_j を包含している $(N_i \supset N_j)$ なら，c_i は c_j の上位の助数詞であると推測できる．例えば，我々のデータベースでは，「店」に呼応する名詞の集合は { スナック, 八百屋, レストラン, ... } であり，これらの名詞は全て「軒」に呼応する名詞の集合にも含まれているた

*北陸先端科学技術大学院大学, Japan Advanced Institute of Science and Technology
†東京工業大学, Tokyo Institute of Technology

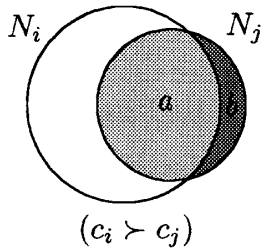

図 1: 名詞集合の包含関係

め，「軒 $\succ$ 店」という関係が推測できる．

そこで，データベースに含まれる全ての助数詞の組について，以下の条件を満たすかどうかを調べ，条件を満たす組については上位-下位関係 $c_i \succ c_j$ が成立すると推測した．

$$|N_i| > |N_j| \tag{1}$$

$$IR(c_i, c_j) > T_{ir} \quad \left(IR(c_i, c_j) \stackrel{def}{=} \frac{|N_i \cap N_j|}{|N_j|} \right) \tag{2}$$

条件 (1) は上位の助数詞ほど多くの名詞と呼応するという考えに基づく．条件 (2) の左辺の包含率 $IR(c_i, c_j)$ は，図 1 における [領域 a+領域 b] に属する名詞数に対する [領域 a] に属する名詞数の割合であり，右辺 T_{ir} はその閾値である．すなわち，たとえ N_i と N_j の間に完全な包含関係が成り立たなくても，包含率 $IR(c_i, c_j)$ がある程度高ければ，2 つの助数詞間に上位-下位関係が成立すると推論する．

閾値 T_{ir} を 0.6 に設定したところ，323 組の助数詞が上位-下位関係にあると推論された．「軒 $\succ$ 店」「枚 $\succ$ 斤」「頭 $\succ$ 蹄」など，直観的に妥当と思われる多くの上位-下位関係が獲得されたことを確認した．獲得された上位-下位関係のより詳細な評価については 4.2 項で述べる．また，上位-下位関係を推測できた助数詞の異なり数は 255 であり，データベース全体に含まれる助数詞の 77%に相当する．

3.2 助数詞オントロジーの構築

3.1 項で獲得した上位-下位関係から助数詞オントロジーを構築する．本研究では，上位の助数詞を親，その下位の助数詞を子とみなして助数詞を単純に連結し，木構造のオントロジーを構築する．すなわち，上位の助数詞を持たない最上位の助数詞を見つけ，それを根ノードとし，それから下位の助数詞を順番に辿って木構造を構築する．このとき，最上位の助数詞は複数あるため，全ての助数詞が連結されて全体で 1 つの木構造ができるわけではなく，複数の木構造が生成される．

また，木構造を構築する際，冗長な上位-下位関係は無視した．冗長な上位-下位関係とは，他の上位-下位関係によって推論可能な関係である．例えば，

$$c_1 \succ c_2, \quad c_2 \succ c_3, \quad c_1 \succ c_3$$

といった 3 つの上位-下位関係が推論されたとき，$c_1 \succ c_3$ は残りの 2 つの関係からも推移律によって推論できる ($c_1 \succ c_2 \succ c_3$) ので冗長である．

表 1: 自動構築されたオントロジーの概要

作成された木構造の数	54
一構造当たりの平均助数詞数	6.0
木構造に含まれる助数詞数の最大値	85
木構造の深さの最大値	3
2 つの助数詞だけからなる木構造の数	24

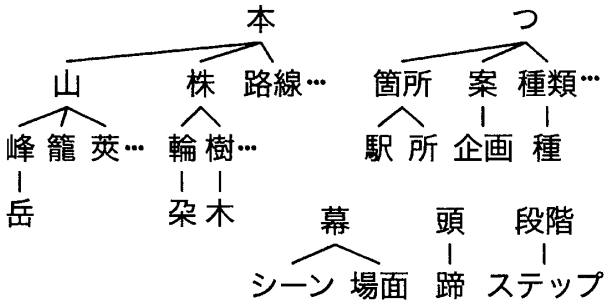

図 2: 自動構築されたオントロジー (一部)

自動構築されたオントロジーの概要を表 1 に，またその一部を図 2 に示す．85 個の助数詞を含む最大のオントロジーは「本」を根とする木構造であった．また，大きい木構造の根には「つ」「個」「枚」など一般的な助数詞が現われることが多かった．一方，全体の 44%に相当する 24 個の木構造は，図 2 の「頭－蹄」「段階－ステップ」のように 2 つの助数詞しか含まない小さな断片であることがわかった．今のところ我々はナイーブな手法しか試していないが，断片化されたオントロジーを何らかの基準によって連結し，全体で 1 つのオントロジーを構築することは今後の重要な課題である．

4 考察

4.1 $N_j \setminus N_i$ に属する名詞の分析

3.1 項で述べた提案手法では，N_i と N_j が完全に包含関係にない場合でも，包含率 $IR(c_i, c_j)$ の値が十分大きければ $c_i \succ c_j$ という関係が成立すると推測する．このとき，差集合 $N_j \setminus N_i$ (図 1 の領域 b) に含まれる名詞に注意する必要がある．関係 $c_i \succ c_j$ は，N_j に属する名詞は全て N_i にも属する (c_i で数えられる) ことを暗に示唆するが，$N_j \setminus N_i$ に属する名詞は c_i で数えることができない例外的な名詞ということになる．ただし，$N_j \setminus N_i$ に属する名詞の中には，データベースを作成する際に参照した辞書 [4] にたまたま記載されていなかっただけで，実際には c_i で数えることができる

表 2: 人手による名詞と助数詞の呼応関係の判定結果

二者の判定が一致した名詞の数	542 (94%)
c_i で数えられる n_k の数	241 (42%)
c_i で数えられない n_k の数	338 (58%)

ものも存在する．もし $N_j \setminus N_i$ に属する名詞の多くが c_i でも数えられることができれば，推測された $c_i \succ c_j$ という関係もより信頼できるといえる．

このような観点から，獲得された $c_i \succ c_j$ について，$N_j \setminus N_i$ に属する名詞を人手でチェックし，それが上位の助数詞 c_i で数えることができるかどうかを判定した．3.1 項で獲得した 255 組の上位-下位関係については，579 個の名詞が $N_j \setminus N_i$ に属することがわかった．次に，これらの名詞 n_k を助数詞 c_i, c_j とともに提示し，n_k が c_i で数えられるかどうかを判定した．判定は著者 2 名が独立して行った．判定が一致しなかった場合は両者の議論によって最終的な判定を決定した．判定のガイドラインを以下に示す．

- 「3 c_i の n_k」(ex. 3 つの玉葱) という表現が妥当なら，n_k は c_i で数えられるとする．
- 上位と下位の助数詞とで数えている単位が違うときは不可と判定する．例えば，(n_k, c_i, c_j) =(玉葱, パック, ネット) のとき，玉葱は「3 パック」とも「3 ネット」とも数えられるが，パックとネットでは数える単位が異なるため，上位-下位関係が成立するとは言い難い．
- 上位と下位の助数詞とで名詞の意味が異なる，あるいは名詞の異なる側面が考慮されているときは不可と判定する．例えば，(n_k, c_i, c_j) =(玉入れ, つ, 回戦) のとき，玉入れは「3 つ」とも「3 回戦」とも数えられるが，「つ」の場合は玉入れという器具を指し，「回戦」の場合は玉入れという競技を指す．このような場合は助数詞に上位-下位関係があるとは言い難い．
- 上位または下位の助数詞が種類を数える助数詞のときは可と判定する．例えば，(n_k, c_i, c_j) =(ソース, つ, 種類) のとき，「3 つ」と「3 種類」では厳密には意味が異なるが，名詞「ソース」の意味自体は同じなので上位の助数詞 c_i でも数えられるとする．

判定の結果を表 2 に示す．二者の判定の一致率は比較的高い値となった．また，判定した名詞の約 4 割が c_i でも数えられることがわかった．これらは元のデータベースでは c_i では数えられないとされていた．このことから，助数詞の上位-下位関係を推測することにより，呼応関係にある名詞と助数詞の組を新たに獲得できる可能性があることがわかる．

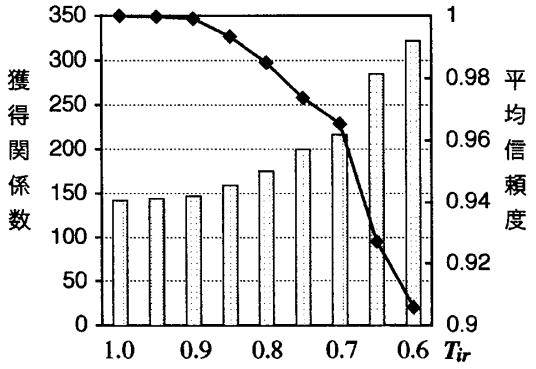

図 3: 包含率の閾値に対する獲得関係数, 信頼度の変化

4.2 上位-下位関係の信頼度

本項では自動獲得された助数詞の上位-下位関係の評価を行う．2 つの助数詞の間に上位-下位関係が成立するとき，上位の助数詞と呼応する名詞は下位の助数詞と呼応する名詞よりも一般的であるので，下位の助数詞と呼応する名詞の多くは上位の助数詞でも数えられるはずである．そこで，下位の助数詞と呼応する名詞のうち，上位の助数詞でも数えられる名詞の割合を求め，それを上位下位関係の信頼度 $R(c_i \succ c_j)$ とする．$R(c_i \succ c_j)$ は式 (3) で求める．

$$R(c_i \succ c_j) = \frac{|N_i \cap N_j| + |NC_{j,i}|}{|N_j|} \quad (3)$$

式 (3) の $NC_{j,i}$ は，N_j の部分集合で，4.1 項で述べた人手による判定で c_i でも数えられるとみなされた名詞の集合である．すなわち，「上位の助数詞でも数えられる名詞」の数を，人手で数えられると判定された名詞の数と $|N_i \cap N_j|$ に属する名詞の数の和としている．

式 (2) における包含率の閾値 T_{ir} を 1 から 0.6 まで変化させたとき，獲得される上位-下位関係の数，およびそれらの信頼度の平均の変化を図 3 に示す．図 3 における棒グラフは関係数，折れ線が平均信頼度の変化を表わす．T_{ir} を下げれば下げるほど，より多くの上位-下位関係が獲得されるが，信頼度は低下することがわかる．ただし，閾値 T_{ir} を 0.6 に設定したときでも，獲得された上位-下位関係の信頼度の平均は 0.91 と比較的高いことがわかる．これは，$N_j \setminus N_i$ に属する名詞の多くが実際には c_i でも数えることができるため，上位-下位関係の信頼度も高く見積られるためである．

4.3 上位-下位関係の推移律の検証

木構造によって表現された助数詞のオントロジーは，子ノードが持つ助数詞の性質が親ノードに継承されると考えられる．すなわち，ある助数詞で数えられる名

表 3: 推移律の妥当性の検証

	信頼度	包含率
比較した (c_a, c_b) の数	55	86
$X(c_a, c_b)$ の平均値	0.93	0.67
A. $X(c_a, c_b) < min_i$	3(0.05)	33(0.38)
B. $min_i \leq X(c_a, c_b) < max_i$	24(0.44)	18(0.21)
C. $X(c_a, c_b) \geq max_i$	28(0.51)	35(0.41)

詞は全てその親または先祖の助数詞でも数えられるとみなせる．すなわち助数詞の上位-下位関係に以下のような推移律が成立することを仮定している．

$c_1 \succ c_2$ かつ $c_2 \succ c_3$ なら，$c_1 \succ c_3$ が成立する

本項では推移律が成立するという仮定がどの程度妥当であるかを検証する．

まず，3.2 項で構築したオントロジーにおいて，先祖-子孫関係にある助数詞の組，すなわち関係 $c_a \overset{*}{\succ} c_b$ が成立する全ての (c_a, c_b) を取り出す．ここで $\overset{*}{\succ}$ は $\succ$ の 0 回以上の適用を表わす．$c_a \overset{*}{\succ} c_b$ は以下のように表現できる．

$$c_1(= c_a) \succ c_2 \succ ... \succ c_n(= c_b) \qquad (4)$$

ここでは推移律に着目しているので，$n \geq 3$，すなわち c_a と c_b のオントロジーにおけるパスの長さは 2 以上であるとする．

推移律によって推測される関係 $c_a \succ c_b$ の信頼度 $R(c_a \succ c_b)$ が，隣接する 2 つの助数詞の組の上位-下位関係 $c_i \succ c_{i+1}$ $(1 \leq i < n)$ よりも同等もしくは高ければ，オントロジーを基にした推移律による上位-下位関係の推論は妥当であるといえる．そこで，先祖-子孫関係にある助数詞の組 (c_a, c_b) と，オントロジーにおける c_a と c_b を結ぶパス上にありかつ隣接している助数詞の組 (c_i, c_{i+1}) とで上位-下位関係の信頼度を比較した．ただし，直接の上位-下位関係にあると推論された助数詞の組に対してしか，4.1 項で述べた $N_j \setminus N_i$ 中の名詞に対する判定を行っていないため，先祖-子孫関係にある全ての助数詞の組に対して信頼度を求めることはできない[1]．一方，式 (2) の包含率 $IR(c_i, c_j)$ は，先祖-子孫関係にある全ての助数詞の組について計算でき，かつ図 3 に示したように信頼度と包含率には正の相関関係があるため，信頼度の良い近似になると考えられる．そこで，信頼度と同様に包含率の比較も行った．

結果を表 3 に示す．表中の $X(c_a, c_b)$ は関係 $c_a \succ c_b$ の信頼度または包含率を表わす．一方，min_i は，c_a と c_b のパス上にありかつ隣接している関係 $c_i \succ c_{i+1}$ の信頼度または包含率の最小値であり，max_i は最大値である．A,B,C は，推移律によって推測される関係 $c_a \succ c_b$ の信頼度 (または包含率) とその間にある隣接関係の信頼度 (または包含率) を比較したとき，それぞれの大小関係に該当する組の数とその全体に対する割合を示している．

信頼度が計算できる 55 組の (c_a, c_b) の平均の信頼度は 0.93 であり，比較的高い値であることがわかった．また，信頼度 $R(c_a \succ c_b)$ が c_a と c_b の間にある隣接関係の信頼度の最小値より小さい場合 (表 3 A.) は約 5%しかなく，また半分以上が最大値よりも大きい (表 3 C.) ことがわかった．このことから，オントロジーによる推移律の推論はある程度妥当であると結論できる．一方，先祖-子孫関係にある 86 組全ての (c_a, c_b) について，同様に包含率を比較すると，約 4 割の組については包含率 $IR(c_a, c_b)$ が c_a と c_b の間にある隣接関係の信頼度の最小値より小さかった．このことは，オントロジー上の推移律の上位-下位関係の推論が妥当でない場合も無視できないほどにはあることを示唆する．

5 おわりに

本研究では，日本語の名詞と助数詞の呼応関係に着目し，助数詞のオントロジーを自動構築する試みについて述べた．予備実験の結果から，有用な助数詞オントロジーを構築できる見込みが得られた．今後の課題として，助数詞オントロジーの規模の拡大などが挙げられる．現在，23%の助数詞については上位-下位関係が推論できず，またいくつかの断片化されたオントロジーが構築されたに過ぎない．多くの助数詞を含み，かつ全体で 1 つの助数詞オントロジーを構築するための手法を探究することが次の目標である．

参考文献

[1] Francis Bond and Kyonghee Paik. Reusing an ontology to generate numeral classifiers. In *Proceedings of the COLING*, pp. 90–96, 2000.

[2] Chu-Ren Huang, Keh-jiann Chen, and Zhao-ming Gao. Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. In *Quantitative and Computational Studies of Chinese Linguistics*, pp. 339–352, 1998.

[3] 飯田朝子. 日本語主要助数詞の意味と用法. PhD thesis, 東京大学, 1999.

[4] 飯田朝子. 数え方の辞典. 小学館, 2004.

[5] 金子孝吉. 助数詞と対象分類 – 文化システムの研究 (3) –. 彦根論叢 第 327 号, pp. 115–140, 2000.

[6] 曹紅荃. 日本語助数詞と中国語量詞の対照分析. Master's thesis, 西安交通大学, 1999.

[1]先祖-子孫関係にあり，かつ直接の上位-下位関係にもある助数詞の組があるとき，直接の上位-下位関係は冗長な関係とみなされてオントロジーの構造には反映されていない．