2007-NL-181 (20)
2007/9/26

呼応する名詞の包含関係に着目した助数詞オントロジーの自動構築と評価

白井 清昭 †　　徳永 健伸 ‡
† 北陸先端科学技術大学院大学　　‡ 東京工業大学

本論文では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手法について述べる．まず，2つの助数詞に対し，それらと呼応関係にある2つの名詞集合に包含関係があれば，助数詞の間に上位-下位関係があると推測する．次に，獲得した上位-下位関係を基に，複数の木構造から構成される助数詞オントロジーを自動的に構築する．提案手法により，呼応関係にある名詞と助数詞の組，9,352組から助数詞オントロジーを自動構築する実験を行った．上位-下位関係がどれだけ信頼できるか，木構造が示唆する上位-下位関係の推移律がどの程度妥当であるか，などといった観点から助数詞オントロジーを評価し，妥当なオントロジーが得られたことを確認した．

Automatic Construction and Evaluation of Ontology for Japanese Classifers based on Subsumption Relations of Agreeing Noun Sets

Kiyoaki SHIRAI†　　Takenobu TOKUNAGA‡
†Japan Advanced Institute of Science and Technology　‡Tokyo Institute of Technology

This paper proposes a method for constructing an ontology of numerative classifiers based on noun-classifier agreement. Superordinate-subordinate relations are first extracted based on subsumption relations of noun sets corresponding to classifiers. An ontology is then automatically constructed using these extracted relations. We conducted an experiment to build an ontology from 9,352 pairs of noun-classifier pairs in Japanese. It was evaluated for the reliability of inferred superordinate- subordinate relations based on the generated ontology and the validity of transitivity of the relations in the ontology. We found the structure of the constructed ontology reasonable.

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，助数詞に関する基礎的な知識として，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基礎的かつ有用な言語資源になりうる．また，日本語教育の面から，教育教材としての活用も期待できる．外国人，特に助数詞をあまり使わない欧州言語の使用者にとって，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．助数詞のオント

2007-NL-181 (20)
2007/9/26

ロジーがあれば，例えば使い方のやさしい一般的な助数詞から教えるといった使用方法が考えられる．

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, 7] 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．

アジア言語の助数詞を対象とした研究もいくつか行われている．Sornlertlamvanich は，タイ語を対象に，呼応関係にある名詞と助数詞の組をコーパスから獲得する方法を提案している [8]．一方，中国語の助数詞を対象とした研究として Huang らによるものがある [2]．Huang らは，名詞と助数詞の呼応関係に基づいて名詞のオントロジーを構築する手法を提案している．これに対し，本研究は名詞と助数詞の呼応関係からオントロジーを構築する点は共通しているが，名詞ではなく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

本節では助数詞オントロジーを構築する手法について述べる．まず，3.1 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，3.2 項では獲得した上位-下位関係を基に木構造のオントロジーを自動構築する手法について述べる．

3.1 上位-下位関係の獲得

本研究では助数詞の上位-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．

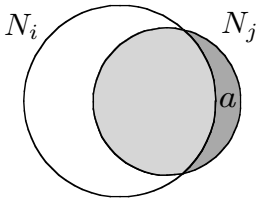

図 1: 名詞集合の包含関係

助数詞の上位-下位関係は名詞と助数詞の呼応関係を基に自動的に獲得する．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．一般に，2 つの助数詞 c_i と c_j があるとき，それらに呼応する N_i と N_j の関係は図 1 のように表わせる．もし，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め,「軒 $\succ$ 店」という関係が推測できる．本研究では，N_i と N_j の包含関係に着目し，以下の 2 つの方法で上位-下位関係にある助数詞の組を獲得する．

3.1.1 包含率を基にした上位-下位関係の獲得

データベースに含まれる全ての助数詞の組について，式 (1) 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 $c_i \succ c_j$ が成立するとみなす．

$$|N_i| > |N_j| \ \text{ かつ } \ IR(c_i, c_j) \geq T_{ir} \qquad (1)$$

$$IR(c_i, c_j) \stackrel{def}{=} \frac{|N_i \cap N_j|}{|N_j|} \qquad (2)$$

式 (1) の $|N_i| > |N_j|$ 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いて設定した条件で

ある．一方，包含率 $IR(c_i, c_j)$ は式 (2) のように定義する．すなわち，包含率とは，下位の助数詞で数えられる名詞のうち，上位の助数詞でも数えることのできるものの割合である．また，T_{ir} はその閾値である．式 (1) の条件は，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みなすことを表わす．

3.1.2 DICE 係数を基にした上位-下位関係の獲得

DICE 係数 (式 (4)) によって 2 つの名詞集合 N_i と N_j の重なりの大きさを測り，DICE 係数が十分大きければ上位-下位関係を獲得する．すなわち，データベースに含まれる全ての助数詞の組から式 (3) の条件を満たすものを求めて上位-下位関係を獲得する．T_{dice} は上位-下位関係が成立するとみなす DICE 係数の閾値である．

$$|N_i| > |N_j| \ \text{ かつ } \ DICE(c_i, c_j) \geq T_{dice} \qquad (3)$$

$$DICE(c_i, c_j) \stackrel{def}{=} \frac{2 \times |N_i \cap N_j|}{|N_i| + |N_j|} \qquad (4)$$

3.1.3 包含率と DICE 係数の比較

包含率の閾値 T_{ir} を 0.6 に設定し，式 (1) を満たす助数詞の組を求めたところ，323 組の助数詞が上位-下位関係として獲得された．一方，DICE 係数の閾値 T_{dice} を 0.3 に設定し，式 (3) を満たす助数詞の組を求めたところ，88 組の助数詞が得られた．「杯 $\succ$ 椀」「枚 $\succ$ 斤」「頭 $\succ$ 蹄」など直観的に妥当と思われる関係が数多く獲得できることがわかった．獲得された上位-下位関係の詳細な評価については 4.1 項で述べるが，ここでは包含率を基準に獲得された関係と DICE 係数を基準に獲得された関係の比較を簡単に行う．

両者に共通して獲得された関係は 55 個であり，獲得される関係に大きな違いがあることがわかった．直観的に正しいと思われる関係は包含率を基準にした方が多かった．特に，DICE 係数で不適切であると思われる関係の多くは $N_i \cap N_j$ の要素数が 1 であった．一方，DICE 係数を基に獲得した関係の方が,「つ」や「個」のような一般的な助数詞が上位となることが少なかった．包含率では上位の助数詞に呼応する名詞集合の大きさ $|N_i|$ は考慮されていない．そのため，多くの名詞と呼応する一般性の高い助数詞は，多くの助数詞に対して呼応する名詞に重なりが生じやすく，包含率が高い助数詞の組が見つかりやすい．一方，DICE 係数では $|N_i|$ も考慮されているため，$|N_i|$ と $|N_j|$ が大きく異なるような関係は獲得されない．「つ」「個」といった一般性の高い助数詞を上位とする上位-下位関係はある程度自明であり，また 3.2 項で述べるように木構造のオントロジーを構築する際にも望ましくない．

このように包含率に基づく手法と DICE 係数に基づく手法には一長一短があるが，包含率を基準に上位-下位関係を獲得した方が有望であるように思われた．以後，上位-下位関係は包含率を基準に獲得するものとする．包含率を基準としたときには一般的な助数詞が上位になりやすいという問題は，後述する関係の「距離」という概念を導入することで対応する．

3.2 助数詞オントロジーの構築

3.1.1 で述べた方法で獲得した上位-下位関係から木構造の助数詞オントロジーを構築する．本研究では，オントロジーを構築する際には上位-下位関係の距離を考慮する．関係の距離とは，上位-下位関係にある 2 つの助数詞の一般性の違いを定量的に評価する尺度である．例えば,「つ $\succ$ 店」という関係では「つ」は「店」よりもはるかに一般的であるが,「軒 $\succ$ 店」という関係における「軒」と「店」の一般性の違いはそれほど大きくはない．つまり,「つ」と「店」の距離は「軒」と「店」の距離よりも大きい．我々は，オントロジーを構築する際，距離が大きい上位-下位関係を反映させるよりも，距離の短かい上位-下位関係をつなげた方が望ましいと考える．先の例では,「つ」が「店」の直接の親になるより,「店」の親が「軒」であり「軒」の親が「つ」であるような構造の方がオントロジーとして自然である．

本研究では関係 $c_i \succ c_j$ の距離 D を式 (5) のように定義する．

$$D(c_i \succ c_j) \stackrel{def}{=} \frac{|N_i|}{|N_j|} \qquad (5)$$

すなわち，上位-下位関係の距離は，上位の助数詞と呼応する名詞の数と下位の助数詞と呼応する名詞の数の比である．我々のデータベースでは，例えば $D(\text{つ} \succ \text{店}) = 19$，$D(\text{軒} \succ \text{店}) = 3.5$ となる．

次に，助数詞オントロジーを自動構築する 4 つの手法について述べる．

3.2.1 手法 1

距離を考慮せず，獲得された上位-下位関係を単純に連結して助数詞オントロジーを構築する．具体的な手続きを以下に示す．

1. 包含率の閾値 T_{ir} を 0.6 とし，式 (1) を満たす上位-下位関係の集合を獲得する．次に，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．
2. 冗長な上位-下位関係を削除する．冗長な上位-下位関係とは，式 (6) のように他の上位-下位関係から推移律によって推論可能な関係と定義する．

$$c_a \succ c_b \text{ は冗長} \quad iff \quad \exists c_m : c_a \succ c_m,\ c_m \succ c_b \quad (6)$$

1. で構築したオントロジーから冗長な上位-下位関係を検出し，それらを全て削除する．

3.2.2 手法 2

まず，包含率が大きくかつ距離が小さい上位-下位関係のみから小さいオントロジーを構築する．次に，包含率と距離の条件を次第に緩和し，オントロジーを段階的に拡張する．以下，i 番目の段階における包含率ならびに距離の閾値を $T_{ir}^{(i)}$，$T_d^{(i)}$ とおく．また，包含率が $T_{ir}^{(i)}$ より大きく[1]，かつ距離 D が $T_d^{(i)}$ 以下である関係の集合を $S(T_{ir}^{(i)}, T_d^{(i)})$ とする．

1. 初期のオントロジーを作成する．$T_{ir}^{(1)} = 0.95$，$T_d^{(1)} = 5$ として $S(T_{ir}^{(1)}, T_d^{(1)})$ を求め，これらの関係を用いて手法 1 の 1. と同じ手続きで木構造を構築する．
2. 親のいない助数詞 c に対して，その親の助数詞 (上位の助数詞) を探す．親のいない助数詞とは，オントロジーに含まれていない助数詞と，構造の根に位置する助数詞の両方を指す．

   まず，閾値を以下のように更新する．

   $$T_{ir}^{(i+1)} = T_{ir}^{(i)} - 0.05, \quad T_d^{(i+1)} = T_d^{(i)} + 70$$

   差集合 $S(T_{ir}^{(i+1)}, T_d^{(i+1)}) \setminus S(T_{ir}^{(i)}, T_d^{(i)})$ 中の関係，すなわち条件を緩和して新たに獲得される上位-下位関係のうち，c が下位となる関係を見つけ，その上位の助数詞を親とみなし，その親子関係をオントロジーに追加する．親の助数詞が複数見つかった場合は対応する全ての関係を追加する．以上の操作を $T_{ir}^{(i)}$ が 0.6，$T_d^{(i)}$ が 495 になるまで繰り返す．
3. 手法 1 の 2. と同じく冗長な関係を削除する．

この手法では，ある助数詞に対する上位の助数詞の候補が複数存在する場合，その全ての上位-下位関係をオントロジーに反映させるのではなく，包含率の高い関係や距離の小さい関係のみを用いてオントロジーを構築する．

3.2.3 手法 3

冗長な関係を削除する際，関係の距離を考慮し，必要に応じて構造の修正を行う．冗長な関係を $c_a \succ c_b$ とし，また c_a から c_b のパス上にある助数詞を c_m とする．もし，(i) $D(c_a \succ c_m)$ が $D(c_m \succ c_b)$ よりも十分大きく，かつ (ii) $D(c_a \succ c_m)$ と $D(c_a \succ c_b)$ に大きな差がないなら，関係 $c_m \succ c_b$ を削除し，c_m と c_b が兄弟となるように構造を修正する (図 2 (a))．相対的に距離が近い上位-下位関係にある 2 つの助数詞は，オントロジー上では親子よりも兄弟として配置した方が直観的に理解しやすいと考えられるからである．また，(ii) の条件は，c_m と c_b を兄弟としたとき，これらと親 (c_a) との距離があまりに不釣合いになることを避けるために設定されている．一方，(i) $D(c_m \succ c_b)$ が $D(c_a \succ c_m)$ よりも十分大きく，かつ (ii) $D(c_m \succ c_b)$ と $D(c_a \succ c_b)$ に大きな差がないなら，関係 $c_a \succ c_m$ を削除し，c_a と c_m が兄弟となるように構造を修正する (図 2 (b))．ただし，本研究では上位の助数詞に下位の助数詞をつなげてオントロジーを構築するので，実際には図 2 (b) の点線で囲まれた 2 つの構造が作成される．ちなみに，手法 2 では図 2(a),(b) いずれの場合も冗長な関係 $c_a \succ c_b$ が無条件に削除される．

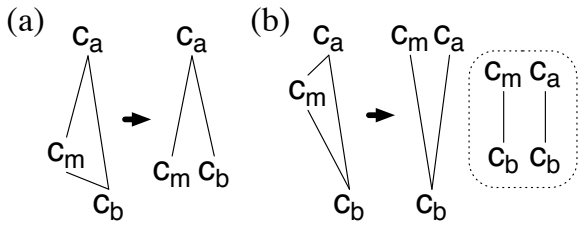

図 2: オントロジーの構造の修正

手法 3 によるオントロジー構築の手続きは以下の通りである．

1. 手法 2 の 1.,2. の手続きでオントロジーを構築する．
2. 冗長な関係 $c_a \succ c_b$ ならびにそれらのパス上にある助数詞 c_m を見つける．

[1] 正確には式 (1) の前半の条件 $|N_i| > |N_j|$ も満たす．

a) 式 (7) の条件を満たすとき，関係 $c_m \succ c_b$ を構造から削除する (図 2(a) の場合).

$$\frac{D(c_a \succ c_m)}{D(c_m \succ c_b)} \geq T_{s1} \text{ かつ } \frac{D(c_a \succ c_b)}{D(c_a \succ c_m)} \leq T_{s2} \quad (7)$$

b) 式 (8) の条件を満たすとき，関係 $c_a \succ c_m$ を構造から削除する (図 2(b) の場合).

$$\frac{D(c_m \succ c_b)}{D(c_a \succ c_m)} \geq T_{s1} \text{ かつ } \frac{D(c_a \succ c_b)}{D(c_m \succ c_b)} \leq T_{s2} \quad (8)$$

c) それ以外のとき，冗長な関係 $c_a \succ c_b$ を構造から削除する.

T_{s1} は 2 つの関係の距離に十分差があることを，T_{s2} は 2 つの関係の距離が同程度であることを評価するための閾値である．本研究では $T_{s1} = 2$, $T_{s2} = 4$ と設定した.

3.2.4 手法 4

より多くの助数詞を含むオントロジーを構築するため，オントロジーに含まれなかった助数詞については包含率の閾値をさらに下げて親の助数詞を 1 つだけ発見し，オントロジーに追加する.

1. 手法 3 によりオントロジーを構築する.
2. $T_{ir}^{(1)}$=0.6 とする.
3. オントロジーに含まれていない助数詞 c について，その親の助数詞 (上位の助数詞) を探す．まず，包含率の閾値を以下のように更新する.

$$T_{ir}^{(i+1)} = T_{ir}^{(i)} - 0.05$$

差集合 $S(T_{ir}^{(i+1)}, \infty) \setminus S(T_{ir}^{(i)}, \infty)$ 中の関係のうち，c が下位となるものを見つける．複数の関係が見つかる場合は距離の小さい関係を 1 つ選択する．見つかった上位-下位関係をオントロジーに追加する．以上の操作を $T_{ir}^{(i)}$ が 0.3 になるまで繰り返す.

4 評価

本節では提案手法の評価を行う．4.1 項では自動獲得された助数詞の上位-下位関係，4.2 項では構造化された助数詞オントロジーの評価について述べる.

4.1 上位-下位関係の評価

4.1.1 $N_j \setminus N_i$ に属する名詞の分析

本研究では，N_i が N_j を完全に包含しない助数詞の組に対しても，包含率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みなす．このとき，差集合 $N_j \setminus N_i$ (図 1 の領域 a) 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，推測された上位-下位関係 $c_i \succ c_j$ もより信頼できるといえる.

このような観点から，上位-下位関係 $c_i \succ c_j$ 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．分析の対象としたのは以下の上位-下位関係である.

- 3.1.1 の手法，すなわち包含率を基に獲得した上位-下位関係．T_{ir} は 0.6 とした．関係の数は 323 である.
- 3.1.2 の手法，すなわち DICE 係数を基に獲得した上位-下位関係．T_{dice} は 0.3 とした．関係の数は 88 である.
- 3.2 項の手法 1 から手法 4 の方法で構築した助数詞オントロジーにおいて，先祖-子孫関係にある助数詞の組の間に成立する上位-下位関係．直接の親子関係を含む．4 つのオントロジーにおける関係の異なり数は 434 である.

上記の 3 種類の関係の和集合における関係の異なり数は 459 である．これらの上位-下位関係に対し，1,373 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人手で判定した．また，1 つの名詞に対して 2 名の被験者が独立して判定を行った．判定のガイドラインについては文献 [6] を参照していただきたい.

表 1: 人手による名詞と助数詞の呼応関係の判定結果

判定者	(s_1,s_2)	(s_3,s_4)	(s_4,s_5)	(s_3,s_5)
判定した名詞数	651	241	241	240
判定の一致率	92.6%	81.7%	84.7%	78.8%

判定の結果を表 1 に示す．判定者の数は 5 名である．s_i は個々の判定者を表わし，s_1 と s_2 は著者 2 名である．二者の判定の一致率は比較的高い値となっ

た．ただし，著者 2 名の判定一致率に比べて他の判定者間の一致率は低い．この理由は詳細には調査していないが，判定のガイドラインの説明が不十分であったことが原因のひとつとして考えられる．

判定が一致しなかった名詞は，判定者 s_1 と s_2 の場合は両者の議論によって最終的な判定を決定した．それ以外の場合は，第三者 (著者のうち 1 名) が最終的な判定を決定した．その結果，全体の約 4 割にあたる 560 個の名詞については 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．このことは，我々のデータベースは呼応関係にある名詞と助数詞を包括的に収録しているわけではないことを示唆する．

4.1.2 上位-下位関係の信頼度

4.1.1 の冒頭で述べたように，上位-下位関係 $c_i \succ c_j$ は，下位の助数詞 c_j で数えられる名詞は上位の助数詞 c_i でも数えられることを示唆する．そこで，下位の助数詞と呼応する名詞のうち，上位の助数詞とも呼応す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 (9) で求める．

$$R(c_i \succ c_j) = \frac{|N_i \cap N_j| + |NC_{j,i}|}{|N_j|} \quad (9)$$

式 (9) の $NC_{j,i}$ は，N_j の部分集合で，4.1.1 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 (1) 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．同様に，式 (3) における DICE 係数の閾値 T_{dice} を 0.7 から 0.3 まで変化させたとき，獲得される上位-下位関係の数と平均信頼度の変化を図 4 に示す．

図 3 から，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．一方，図 4 から，DICE 係数を基準に上位-下位関係を獲得したときも，閾値 T_{dice} を下げれば獲得される関係数は増える一方，平均信頼度は低下することがわかる．ただし，獲得される関係の数は包含率を基準にしたときと比べて 3 割程度であり，平均信頼度も低い．

4.2 助数詞オントロジーの評価

4.2.1 自動構築されたオントロジーの概要

3.2 項で述べた手法 1 から手法 4 によって自動構築されたオントロジーの概要を表 2 に示す．それぞれの手法によって 50 個程度の木構造が構築されたことがわかる．一構造当たりの助数詞は 5〜6 程度であるが，2 つの助数詞しか含まない小さい断片も 35%から 50%近くあることがわかる．現段階では，1 つの大きな木構造ではなく，小さい木構造の集合として表現された助数詞オントロジーが得られている．

手法 1 と手法 2 を比べると，手法 2 の方が上位-下位関係の平均信頼度が高い．一方，構造の数や上位-下位関係の数を比べると，手法 2 は手法 1 と比べてオントロジーの規模が小さいといえる．これは，手法 2 は上位-下位関係を獲得するための包含率や距離の条件を段階に緩和し，相対的に包含率の低い関係や距離の大きい関係を構造に反映させていないためである．

手法 2 と手法 3 の違いは，手法 3 は冗長な関係を削除する際に距離を考慮して構造の修正を行う (3.2.3 の 2. の手続き) 点にある．手法 3 において構造を修正した例を図 5 に挙げる．

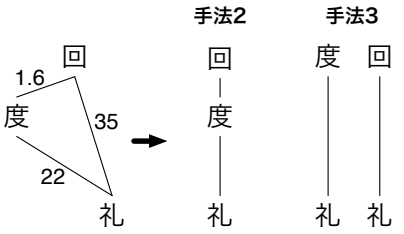

図 5: 構造の修正例

図 5 中の数字は関係の距離を表わす.「礼」は挨拶やお辞儀の回数を数える助数詞である．手法 2 では冗長な関係「回 $\succ$ 礼」が削除され，手法 3 では距離の近い「回 $\succ$ 度」が削除される．回数を数える一般的な「回」と「度」の間に上位-下位関係が成立するとする手法 2 よりも，特殊な助数詞「礼」が「回」や「度」の子に位置する手法 3 の方が適切であると思わ

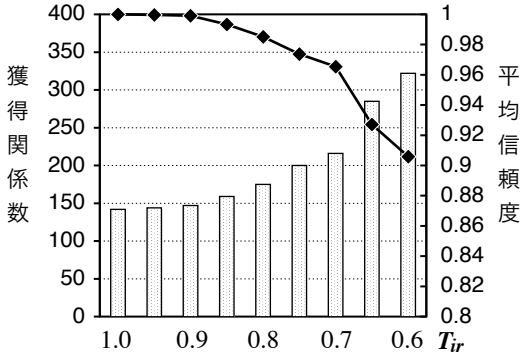

図 3: 包含率に基づく上位-下位関係の評価

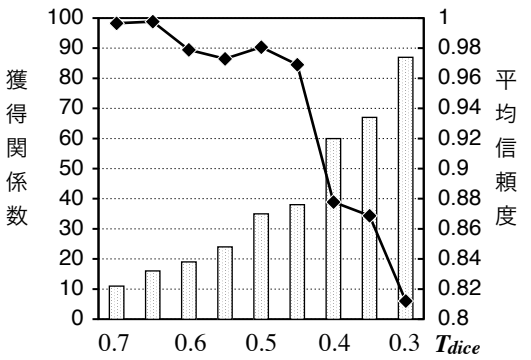

図 4: DICE 係数に基づく上位-下位関係の評価

表 2: 自動構築されたオントロジーの概要

		手法 1	手法 2	手法 3	手法 4
[木構造]	作成された木構造の数	54	47	53	57
	一つの木構造に含まれる助数詞数の平均	6.3	5.8	5.3	5.2
	一つの木構造に含まれる助数詞数の最大値	85	53	53	66
	深さの最大値	3	3	3	3
	2 つの助数詞だけからなる木構造の割合	44%	43%	49%	35%
[上位-下位関係]	関係数	274	225	225	300
	平均信頼度	0.903	0.922	0.920	0.860
[助数詞]	助数詞の異なり数	255	242	242	321
	全助数詞数に対する割合	77%	73%	73%	97%

れる．手法 3 で実際に修正が適用されたのは 7 箇所であった．また，全て図 2 (b) に該当する修正であった．ただし，修正の箇所が少ないため，表 2 における手法 2 と手法 3 の違いはそれほど大きくない．手法 3 の方が木構造の数が多いのは，図 5 のように断片的な構造が新たに生成されるためである．

手法 4 では，手法 1〜手法 3 と比べて規模が大きいオントロジーが構築されていることがわかる．特に，オントロジーに含まれる助数詞の割合が 77%から 97%にまで改善されている．これは，包含率の低い関係を使って助数詞の追加を行った (3.2.4 の 3. の手続き) ためである．一方，上位-下位関係の平均信頼度は 0.86 に低下している．当然ながら，オントロジーの規模と信頼性 (上位-下位関係の信頼度) はトレードオフの関係にある．

手法 4 で作成されたオントロジーの一部を図 6 に示す．66 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．

4.2.2 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詞は全てその親または先祖の助数詞でも数

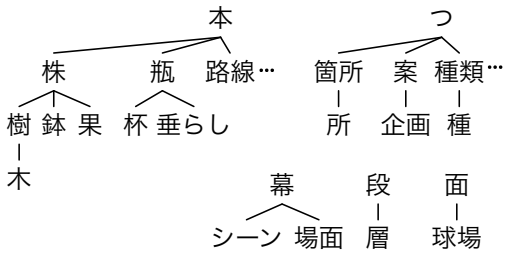

図 6: 手法 4 で自動構築されたオントロジー (一部)

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

ここ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ことにより，自動構築されたオントロジーの評価を行う．

まず，自動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 $c_a \overset{*}{\succ} c_d$ が成立する全ての (c_a, c_d) を取り出す．ここで $\overset{*}{\succ}$ は $\succ$ の 0 回以上の適用を表わす．$c_a \overset{*}{\succ} c_d$ は以下のように表現できる．

$$c_1(= c_a) \succ c_2 \succ ... \succ c_n(= c_d) \qquad (10)$$

表 3: 推移律の妥当性の検証

	手法 1	手法 2	手法 3	手法 4
先祖-子孫関係にある (c_a, c_d) の数	86	54	41	53
$R(c_a, c_d)$ の平均値	0.77	0.78	0.85	0.78
A. $R(c_a, c_d) < \min_i$	24 (28%)	14 (26%)	8(20%)	10 (18%)
B. $\min_i \leq R(c_a, c_d) < \max_i$	27 (31%)	22 (41%)	18(44%)	30 (54%)
C. $R(c_a, c_d) \geq \max_i$	35 (41%)	18 (33%)	15(37%)	16 (29%)

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_d 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 $c_a \succ c_d$ 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，構造化されたオントロジーに基づく推移律による上位-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_d) と，オントロジーにおける c_a と c_d 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．

結果を表 3 に示す．表中の $\min_i$ は，c_a と c_d 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 $c_a \succ c_d$ の信頼度とその間にある隣接関係の信頼度を比較したとき，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．

各手法で構築されたオントロジーにおいて，推移律によって推測される上位-下位関係の信頼度 $R(c_a, c_d)$ の平均は 0.77 から 0.85 であり，表 2 に示した直接の上位-下位関係の平均信頼度よりは低い．しかし，$R(c_a, c_d)$ が c_a から c_d のパス上に隣接している上位-下位関係の信頼度よりも小さい場合 (表 3 の A.) は 2 割から 3 割程度である．このことから，自動構築したオントロジーにおける推移律はある程度成立すると言える．また，手法 3 は手法 2 と比べて，$R(c_a, c_d)$ の平均信頼度，ケース A. に分類される割合がともに大きく改善されている．手法 3 による構造の修正は 7 箇所とわずかではあるが，これにより推移律が成り立たなくなるような不適切な上位-下位関係が削除され，適切に構造が修正されたといえる．

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．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．

最後に今後の課題について述べる．本論文では，助数詞の上位-下位関係を助数詞と呼応する名詞集合の包含関係に着目して推論しているが，これとは別の観点から上位-下位関係を獲得することを検討している．具体的には，2 つの助数詞と呼応する 2 つの名詞集合において，一方の集合に含まれる名詞と他方の集合に含まれる名詞の組に着目し，多くの名詞の組に対して上位-下位関係が成立すれば助数詞の間にも上位-下位関係が成立するとみなす．本論文で提案した包含関係を基準とする場合と名詞間の上位-下位関係を基準とする場合とで，獲得される助数詞の上位-下位関係やオントロジーの構造にどのような違いが生じるかについても調査したい．

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] 白井清昭, 徳永健伸. 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. 情報科学技術フォーラム FIT2007, E-009, pp. 147–150, 2007.

[7] 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.

[8] Virach Sornlertlamvanich, Wantanee Pantachat, and Surapant Meknavin. Classifier assignment by corpus-based approach. In *Proceedings of the COLING*, pp. 556–561, 1994.