辞書に基づくオントロジー生成学習

鈴木 敏 (satoshi@cslab.kecl.ntt.co.jp)
NTT コミュニケーション科学基礎研究所

辞書を基にした単語オントロジーの自動生成手法を提案する。提案するのは、単語の定義文を基に上位語候補を抽出し、最適な木構造を推定する手法である。本手法の特徴は、語義文の再帰展開手法と、弱い相関から木構造を取り出すデータマイニング手法の組合せにある。再帰展開手法は定義文中に明示されていない単語をも候補とすることを可能とし、上位語と相関のある情報を作り出す。また、データマイニング手法は前記の上位語相関情報を利用して単語間の最適な上下関係を推定することができる。この結果、辞書のみから全く新しい単語の上下関係を表す木構造を生み出すことが可能となる。本論文では、これらの手法の詳細を示し、単語オントロジーの生成結果について詳細を述べる。

1　はじめに

言語処理のリソースとして、オントロジーは重要なデータの一つである．例えば，日本語のオントロジーとしては日本語語彙大系 (池原，宮崎，白井，横尾，中岩，小倉，大山，林 1997) 等が挙げられるが、これは翻訳をはじめとして様々な言語処理に利用されている。

このオントロジーは人手により編集されたものであるが，言葉が日々変化するものであることを考えると頻繁な更新が必要であり，手作業によるオントロジー構築には自ずと限界がある．従って，このようなオントロジーの構築は自動化されることが望まれる。

オントロジーは単語の意味的関連性を表すものであるが、単語間の意味関係は一つではなく，従って，オントロジーもその関連性によって様々に変化すべきものである．この点を踏まえて自動化を考えると，Snow らの提案するような単語を既存のオントロジーに追加する手法ではなく (Snow, Jurafsky, and Ng 2006)，木構造をゼロから作り出せるような手法が必要となる．本論文では、この点を主眼として木構造構築のための学習手法の提案を行う．

ところで、語義文中に上位語が含まれているという仮定の下で、辞書から単語の上位語と相関のある情報を取り出す手法がある (鈴木 2007)．本稿では、この手法により取り出された相関情報を用いたオントロジーの構築を試みる．

以下，鈴木により提案された、語義文の拡張による上位語相関情報の抽出手法を説明し，次いで，その情報を用いたオントロジー構築のための学習手法を示す．また、実験では約 44000 語の名詞によるオントロジー構築を試み、日本語語彙大系との比較による評価結果を示す．

2　再帰的語義展開と上位語情報

本節では、辞書の特性を利用する再帰的語義展開手法 (Suzuki 2003; 鈴木 2005) およびその結果得られた情報と上位語との相関 (鈴木 2007) について概要を示す．

2.1　再帰的語義展開

再帰的語義展開の基本的な考え方は，語義文中の単語をその語義文により再帰展開し，より多くの単語からなる語義文を作成するということである。このとき問題となるのは，このような展開が無限に続いてしまうことである。この場合，展開された語義文中の単語数は無限になり，頻度計算は一般に不可能になる．しかしながら，この展開の過程に確率的要素を組み込むことにより，無限に展開された語義文からの単語頻度計算を可能にする方法がある．

語義文の展開を行なう毎に，一定の割合でその影響が小さくなるとすれば，無限に展開された語義文の影響力は元の語義文に比べて微小になる．従って，語義文の影響力は語義展開の回数に比例する等比数列として表すことができる．同時に，その合計を無限級数として計算すると必ず有限な値となる．これにより，語義文の集合体中の単語頻度も有限になり，計算可能となる．これらを確率モデルに置き換えることで，無限の展開を含めた語義文の集合体から単語の出現確率を取り出すことが可能になり，拡張された語義文として再定義することができる．以下，n 回展開された語義文を n 次の語義文と呼ぶ．ただし，辞書中の語義文を 1 次語義文とする．

まず，見出語 w_i と語義文中の単語 w_j の関係は $P(w'_j|w_i)$ と表すものとする．この表記を用いると，全ての見出語に関して，語義文中の単語の出現確率は

$$A = \begin{bmatrix} P(w'_1|w_1) & \cdots & \cdots & P(w'_1|w_m) \\ P(w'_2|w_1) & \ddots & & \\ \vdots & & \ddots & \\ P(w'_m|w_1) & & & P(w'_m|w_m) \end{bmatrix} \quad (1)$$

の列ベクトルとして表される．ここで，m は辞書中の見出語の数である．行列 A の各要素 $P(w'_j|w_i)$ は $N(w)$ を語義文中の単語頻度として，

$$P(w'_j|w_i) = \frac{N(w'_j)}{\sum_{all\ k} N(w'_k)} \quad (2)$$

である．全ての列ベクトルは，要素の合計が 1 であり，確率表現となっていることがわかる．

目的とする単語の拡張語義文は，上記の手法で求められた確率頻度を 1 次から無限次まで確率的に加えた確率ベクトルで表される。単語の頻度は 2 次の語義文では A^2、3 次の語義文では A^3 で表されるため，目的とする拡張語義文の行列表記を C とすると，C は

$$C = P_1A + P_2A^2 + \cdots + P_nA^n + \cdots, \quad (3)$$

表 1: 辞書の語義文と拡張語義文の比較

見出語：通信

通信	4	人	0.0473
人	1	有線	0.0430
有線	1	無線	0.0429
無線	1	自分	0.0424
自分	1	連絡	0.0423
連絡	1	状況	0.0419
状況	1	知らせ	0.0413
知らせ	1	互い	0.0408
互い	1	相手	0.0407
相手	1	意志	0.0397
意志	1	宇宙	0.0393
宇宙	1	思想	0.0387
思想	1	デジタル	0.0383
デジタル	1	通信	0.0373
計 14 語		計 10609 語	

である．ただし，P_n は $\sum_{n=1}^{\infty} P_n = 1$ を満たす．確率ベクトルの確率的総和もまた確率ベクトルである．

語義展開の度に P_n が一定の割合 a で減少すると仮定すると，C は無限級数の計算から

$$(I - aA)C = (1 - a)A \tag{4}$$

を満たし、この線型方程式を解くことにより C の解を求めることができる。

C の (j, i) 要素を $P(w_j^*|w_i)$ と書くと，w^* は展開された語義文集合の中の単語を意味することになる．すなわち，$P(w_j^*|w_i)$ は拡張語義文中の単語の確率頻度である．

2.2 拡張語義文

上記の手法を実際に国語辞典 (金田一，池田 1988) に適用した結果を以下に示す．前処理として，扱う単語を一般名詞とサ変名詞に限定（形態素解析は茶筅 (松本，北内，山下，平野，松田，高岡，浅原 2000) を利用）し，その結果，44,050 語の見出語と，平均約 7 語の 1 次語義文を得た．以下の計算は，$a = 0.5$ で行っている．

まず，式 (1)(2) から確率行列 A を計算した．これはスパースな 44050 次元の正方行列である．次に式 (4) から線形学習法により C を求めた．このときの有効桁は 10^{-6} までとした．学習の結果，十分な収束を得られなかった語を除いて，43,616 語の確率ベクトルを得た．これらのベクトルの非ゼロの値を持つ次元数は平均約 10,000 であった．

表 1 に辞書の語義文と拡張語義文との比較を示す。見出語「通信」に対して、辞書の語義文では「通信」が頻度 4 で最も高い確率を示してるが、拡張語義文では「人」が最も頻度の高い単語となっている。また、辞書の語義文で現れていた「デジタル」は拡張語義文では 14 位以内から消え、辞書の語義文には現れていなかった「文字」が拡張語義文では 10 位の頻度で現れている。また、語義文中の文字数も、辞書の語義文では 14 語だったものが、拡張語義文では 10,609 語に大幅に増えている。

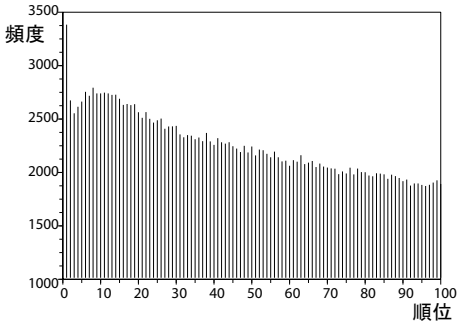

図 1: 出現頻度順位と上位語の関係

2.3 上位語情報としての評価

見出語の語義文中に上位語があるとすれば，上位語 j は見出語 i を説明するために非常に重要な単語であるため，その出現確率 $P(w_j^*|w_i)$ は高いことが予想される．従って，見出語 i の上位語は，その確率ベクトルの要素の中から出現確率 $P(w_j^*|w_i)$ が高い順に尤もらしいと考えることができる．これを証明するために，日本語語彙大系 (池原他 1997) を正解データとした検証実験を行った．対象としたのは，上記 43,616 語のうち，日本語語彙大系に記載のある 39,771 語である．

まず，対象となる全単語に関して，その確率ベクトル中の要素を出現確率が高い順に並べ替え，各要素が見出語の上位語として日本語語彙大系に記載されているかどうかを統計的に調べた．図 1 にその結果を示す．横軸に単語の拡張語義文での出現しやすさの順位を，縦軸に日本語語彙大系との一致度数 (39,771 語中，当該単語が日本語語彙大系での上位語と一致した数) を示してある．図は 100 位までの統計量を示している．図からは順位が高い程，日本語語彙大系に一致する割合が高いことがわかる。この結果は，拡張語義文中の出現確率が高い程，その単語が上位語である可能性が高いことを示唆している。

因みに，図 1 で出現確率 2 位から 6 位の間で大きく値が下がっているのは，同義語が集中するためであると考えられる．

ところで，出現確率最大の値でも正解率は 8.8%であり，決して良い値ではない．しかしながら，不正解となる単語を実際に見ると，上位語として必ずしも間違いとは言えないような単語がいくつも見受けられる．これは，日本語語彙大系が上位語全てを網羅しているわけではなく，見方によって様々なオントロジーが構築できることを示している．

3 オントロジーの構築

3.1 簡易構築

2.2 節で得られた上位語情報を利用してオントロジーを構築する最も簡単な方法は、拡張定義文中の出現頻度が最大の語を当該見出語の上位語とみなして木構造を構築するという方法である。この手法により得られた木構造の例を図 2 に示す．

この図から、直接の上下関係を持つ単語間にはそれぞれ強い関係を見出すことができる。しかしながら、それらの関係

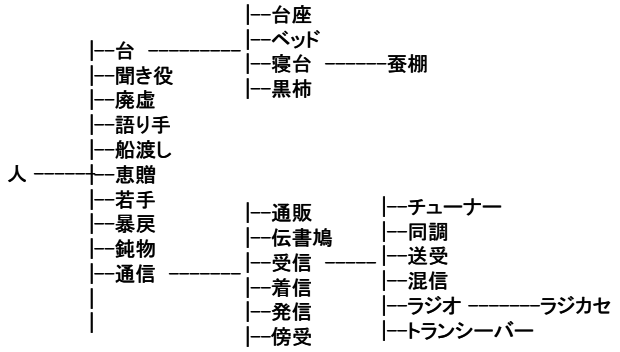

図 2: オントロジーの簡易構築例

は一貫性を保っていない．このため，距離の離れた上下関係を持つ単語間には上下関係を見出し難いものが散見される。この結果から、オントロジー構築のためには距離の離れた上下関係を考慮して構築する必要があることがわかる．

Snow らは離れた単語間の関係を学習する手法を提案しているが (Snow et al. 2006)、この手法は既成のオントロジーに単語を追加するためには有効であるが，本稿で目的とするオントロジーの骨格を生成するための有効な手段とはいえない．従って，オントロジーを構築するための新しい学習手法を試みた．

3.2 学習モデル

拡張語義文中の各単語は、当該見出語を構成する要素であるとし、オントロジー上で見出語自身とその上位に存在する構成要素のみが、その見出語の構成要素として有効であると仮定する。即ち、オントロジー上に単語 A の上位語として単語 B, C のみが存在する場合、A の拡張語義文は A, B, C のみで構成され、語義文中の他の単語は無視されると考える (図 3 参照)．単語の確率頻度 (2.1 節参照) で書くと、

$$P'(w_A) = P(w_A^*|w_A) + P(w_B^*|w_A) + P(w_C^*|w_A) \quad (5)$$

となる。以下，この $P'(w)$ を意味的再現率と呼ぶことにする。

上記仮定の下で、あるオントロジー T が存在する場合、T が存在する確からしさは全単語の意味的再現率の積、即ち、

$$P(T) = \prod_{w \in all} P'(w) \quad (6)$$

により与えられる。つまり、全ての単語が意味を持つ状態にある時がオントロジーの存在が最も安定している状態であると考える。この対数をとれば,

$$\begin{aligned} L(T) &= \log P(T) = \sum_{w \in all} \log P'(w) \\ &= \sum_{w \in all} \log \sum_{w' \in hyper\ w} P(w'|w) \end{aligned} \quad (7)$$

となる。ただし、$w' \in hyper w$ は w の全上位語を意味する。このオントロジーの確からしさ $L(T)$ を最大化することにより、前述の仮定の下での最適な単語オントロジーが取り出せる。

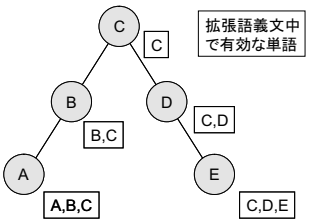

図 3: オントロジーと有効な語義文との関係

3.3 計算機実験

本論文では、計算コストを抑えるため，前述の最適化アルゴリズムを各単語の意味的再現率の最大化と、その組合せとしての全体の最適化との 2 段階に分離し、

1. $P'(w) = \sum_{w' \in hyper\ w} P(w'|w)$ の最大化
2. $L(T) = \sum_{w \in all} \log P'(w)$ の最大化

を交互に行うことで、近似的に最適化を行った。即ち，はじめに単語の意味的再現率を最大にする直接上位語を捜し出し，次に木構造全体としてのオントロジーの確からしさが上昇するかを判断する。これが上昇するのであれば、この単語の直接上位語を当該語に置き換え，上昇しなければ元のままとし、次の単語を検証する。これを全単語で上位語が変化しなくなるまで繰り返した．

オントロジー上での各単語の上位語は 1 語に限定し、自身を上位語とすることも許可する。この場合，当該単語は独立した木構造の最上位語となる。学習のための初期値は、全ての単語が自身を上位語としている状態とした。

オントロジー生成の対象となる単語は、2.2 節で確率ベクトルが得られた名詞 43,616 語である。また、$P(w'|w)$ は見出語 w の拡張語義文中の確率頻度である．以下にその結果を示す．

学習により生成されたオントロジーは「語」を最上位概念とする一つの大きな木構造と、およそ 150 の小さな木構造の組合せとなった。巨大な木構造の大まかな形状は、表 2 に示すように、「語」の直下の概念が多く、下層に行くに従い少なくなっている。6 層目では 1 概念である。この 6 層目の概念は「一つ」で、その配下には再び多くの概念が広がっている。最も深い概念は 119 層であった。また、学習後のオントロジーの特徴の一つとして、代表的な単語の直下に非常に多くの単語がぶら下がっていることも挙げられる。

これら結果から、学習により単語間の上位下位の関係が大きく変化したことがわかる。例えば,「通信」の直接上位語は、図 2 に示される簡易構築では「人」であったのに対し、学習後には「表現」となっている。また、「通信」の直接下位語は「電報」と「活動」であり、「電報」の下位に「着信」「発信」等がぶら下がっている (図 4 参照)。

3.4 評価

次に、生成されたオントロジーの精度を調べた。これは、日本語語彙大系のカテゴリー間の上下関係との一致度を測ることにより実現した。日本語語彙大系には約 3,000 の意味カ

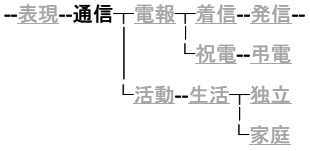

図 4: 学習後の木構造例

表 2: 学習後の木構造の概観

レイヤー	概念数	語
1	1	語
2	1942	
3	160	
4	38	
5	21	
6	1	一つ
30	3	
60	25	
90	1635	
118	4	

テゴリーが与えられており，意味カテゴリー間の上下関係がオントロジーとして与えられている。これら意味カテゴリーと前記実験で用いた名詞 43,616 語との共通の単語は 988 語であった。これらの単語に関して，日本語語彙大系において上下関係が与えられている単語の組合せ 3562 組の再現率を調べた。ここでは、上下関係が入れ替わっている場合は不正解とし、距離が離れていても上下関係が正しいものは正解とした．

結果は再現率 45.5%であった。また、初期状態での再現率は上位語を自身としているため 0%、簡易構築 (3.1 参照) では 7.7%であった。この評価手法では上下の階層が深い程、良い結果を出す傾向があるが、木構造の最深部でも上位語数は 120 程度で、全単語数 43,616 語に比較して微小であり，木構造が深くなった点を考慮しても学習による効果が大きいことがわかる。

この結果は、計算によって得られた上位語情報を、学習によって木構造に組み上げることができるということを示している。また同時に、得られた上位語情報の有効性を示すものでもある．

ただし、主観的評価では、得られた木構造は、簡易構築、学習による構築共に玉石混合といった感じで、実用性という観点からすれば、今回得られた結果はまだ実用に耐え得るレベルのものではない。このことは、前項に示した木構造の形状からもわかり，必要以上に木構造を大きくする傾向がこの学習手法にはある。また、実装アルゴリズムは局所的な最適解に陥りやすいという問題を抱えている。

しかしながら，今回用いた学習手法はあくまでも簡易的なものであり、局所的な最適解しか得られていない点を考慮すれば、今後の更なる精度の向上も期待できる。また、今回提案した手法以外の最適化手法を導入することも、もちろん可能である．

4 おわりに

本論文では辞書の語義文から上位語情報を取り出し、これを基に単語の意味的上下関係を表すオントロジーの生成を試みた。結果は，学習の効果としてオントロジーの精度の向上は示せたが，オントロジーの精度としては実用レベルにはまだ届いていないものであった．

今後の課題としては、第一に、新しい学習手法を含めて、より高精度の学習手法を考案することが挙げられる．今回の結果のように、一つの巨大な木構造に単語を集中させるのではなく、複数の木構造に分散したオントロジーを可能にするような学習手法も検討したい．

また、上位語情報の高精度化も大きな課題である．そのアプローチとして、より多くの辞書を利用すること、学習データに “is-a” 構造から取り出したデータを組み込むこと等も考えられる。

更に、今回の実験では語義の曖昧性については考慮せず、1 表記につき 1 語義として実験を行ったが，語義文中の語義曖昧性を排除した辞書 (例えば lexeed(笠原，佐藤，田中，藤田，金杉，天野 2004)) を利用することにより、語義レベルのオントロジー生成を行うことも検討中である．

参考文献

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). “Semantic Taxonomy Induction from Heterogenous Evidence.” In *Proceedings of 44th Annual Meeting of the ACL*, pp. 801–808. Association for Computational Linguistics, ACL.

Suzuki, S. (2003). “Probabilistic Word Vector and Similarity based on Dictionaries.” *Computational Linguistics and Intelligent Text Processing Lecture Notes in Computer Science (In proceedings of CICLing2003)*, **N 2588**, pp. 564–574.

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

金田一春彦，池田弥三朗 (1988). 学研国語大辞典第二版. 学習研究社.

鈴木敏 (2005). “辞書に基づく単語の再帰的語義展開.” 情報処理学会論文誌, **46** (2), pp. 624–630.

鈴木敏 (2007). “辞書を用いたオントロジー自動生成　－上位語候補の自動抽出.” 言語処理学会　第１３回年次大会　併設ワークショップW１　言語オントロジーの構築・連携・利用, pp. 47–50.

松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸 (2000). “日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書.” *http://chasen.aist-nara.ac.jp/*.

笠原要，佐藤浩史，田中貴秋，藤田早苗，金杉友子，天野成昭 (2004). “「基本語意味データベース:Lexeed」の構築 (辞書, コーパス).” 情報処理学会研究報告. 自然言語処理研究会報告, **2004** (1), pp. 75–82.