

MDL 原理に基づく辞書定義文の圧縮と共通性の発見

土屋 雅稔

黒橋 禎夫

京都大学大学院情報学研究科

〒606-8501 京都市左京区吉田本町

tsuchiya@kuee.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

要旨

本論文は、計算機によって意味を取り扱うことを目的として、辞書定義文中に含まれる種々の特徴を分類し、単語間の相違点を自動的に取り出す方法を提案する。辞書定義文には、見出し語を説明するための種々の特徴が含まれているが、これらは大別して、共通特徴、共通変数および個別特徴の3つに分類することができる。この操作を自動的に実行するため、既存の自然言語処理技術の応用によって辞書をグラフ集合に変換し、そのグラフ集合をMDL原理に基づいて圧縮するアルゴリズムを考案した。実際に、小学生用の辞書を対象として提案手法を適用し、我々の言語直感に概ね適合するような特徴の発見が可能であることを示した。

Compression of Description Sentences and Discovery of Common Features based on MDL Principle

Masatoshi Tsuchiya,

Sadao Kurohashi

Graduate School of Informatics, Kyoto University

Yoshida-honmachi, Sakyo, Kyoto, 606-8501 Japan

tsuchiya@kuee.kyoto-u.ac.jp, kuro@i.kyoto-u.ac.jp

Abstract

The paper proposes an automatic method of feature extraction of an ordinary dictionary in order to enable machine to handle meanings. We employed the MDL principle as criteria to compress a dictionary, which leads to distinguish common features, common slots and characteristics of definitions. The experiment on a Japanese children dictionary is reported, showing the effectiveness of our method.

1 はじめに

計算機による知的処理を実現する上での最大の課題として、計算機上でどのように単語の意味を取り扱うか、という問題がある。この点について、多くの研究は、シソーラスに記述された意味分類を参照し、その意味分類によって単語の意味が表現されていると見なしている。しかし、ほとんどの既存のシソーラスは、類義語間の意味・特徴の相違点や、上位語と下位語がどのように関係しているか、などの情報が欠落しているため、意味の記述として不十分な点が多い。

自然言語で記述された通常の辞書には、そのような情報も豊富に含まれている。しかし、自然言語のテキストは非常に自由度が高く、同じ事柄に対する記述でも様々な記述が可能であるため、そのままの形式で計算機から利用することは難しい。それに対して、人工的な知識表現言語を設計し、それを用いて人手によって同等の辞書を記述するという方法を考えることもできるが、その場合には、知識表現言語の記述力、人手による作業のコスト、一貫性などが問題となる。

本論文では、辞書を対象として意味の類似した一群の単語と、それらの単語間の相違点を取り出す方法を提案する。この方法は自然言語テキストから知識ベースを自動構築する方法の第一歩であると位置づけることができる。

2 定義文の特徴

まず、辞書定義文の性質についての考察を通じて、どのような共通性が存在し、どのような共通部分を取り出されるべきかについて検討する。

例として、花の咲く植物についての3つの定義文について考える。

アブラナ 春、黄色い花がさく。

瓢箪 夏に白い花がさく。

ねむの木 夏、赤い花が房のようにさく。

これらの定義文には「花がさく」という表現が共通して現れているが、この部分は「これらの

見出し語に共通する特徴」を表現していると考えることができる。逆に考えると、この特徴的な表現部分が存在するからこそ、これらの見出し語に類似性が見い出されたとも言える。本論文では、このような特徴を**共通特徴**と呼ぶ。

次に、「黄色い」「白い」のように花の色を表している部分や、花の咲く季節を記述している部分が見つかる。これらの部分は、「共通特徴 (= 花が咲くこと) を説明するための共通の属性」を表現している部分と考えられる。このような部分を**共通変数**と呼ぶことにする。

最後に残った部分は、見出し語毎に個別の特徴を記述した部分と見なすことができる。この部分を**個別特徴**と呼ぶ。

しかし、共通変数と個別特徴の違いは、非常に曖昧である。例えば、「ねむの木」の定義文中に現れる「房のように」という表現は、他の定義文には現れないという観点からは個別特徴と考えられるが、花の咲く様子について説明している部分であると考えれば、共通変数的であるとも言える。この点について、人手に基づく知識ベースの設計者は個人の言語的直感に頼って対処してきたが、本論文の提案する手法では、この問題を情報理論的なアプローチで解決することができる。

3 計算機による特徴抽出

この章では、定義文からの特徴抽出を計算機によって自動的に行う方法について検討する。本論文の提案手法の基本的な手順と仮定は次の通り。

個々の定義文を構文解析することによって、辞書をグラフ集合に変換する。得られたグラフ集合を効率よく圧縮すると、その結果が辞書定義文の特徴的な部分を表す。

以下、細部について順に説明する。

3.1 辞書のグラフ表現

日本語の文は、一般的に、文節を単位とする依存構造によって説明される。この時、個々の

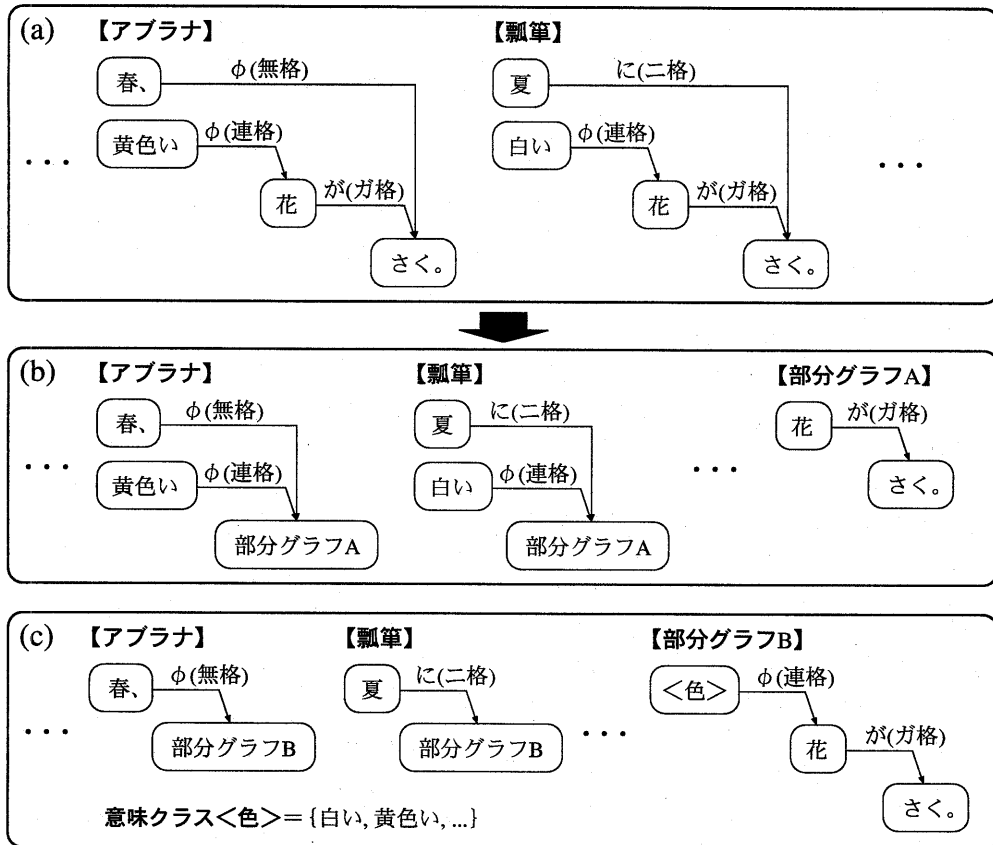


図 1: 辞書の圧縮

文節に含まれる自立語が、その文節の主要な意味を表現し、残りの付属語部分および自立語の活用形が、その文節の他文節に対する係り方を表している。このように考えると、日本語文の構造は、節点のラベルが自立語、辺のタイプが係り受けの性質に対応するようなラベル付き有向グラフと見なすことができる。例えば、「アブラナ」「瓢箪」についての定義文の構造は、図 1-a のようなグラフに変換される。

3.2 グラフ表現と特徴抽出

次に、取り出されるべき個々の特徴に該当する部分が、グラフ表現としては、どのような部分に該当するかを検討する。

共通特徴を表す部分は、その性質から、幾つかの定義文に共通して現れる部分グラフに対応

すると考えられる。これは図 1-a にも明らかに示されているように、「アブラナ」「瓢箪」の定義文の共通特徴である「花がさく」という部分は、2つのグラフに共通して現れる部分グラフとして表現されている。

共通変数に該当する部分について考えると、図 1-a において、「黄色い」「白い」などの共通変数にあたる部分は、共通特徴に対応する部分グラフに対して、係り受けを持つ節点となっている。共通変数は、共通特徴を説明するための共通の属性を表現している部分であるから、共通特徴部分に対する修飾語として現れるのは、極めて自然だと考えられる。そこで、共通変数に対応する節点のラベルを要素とする集合を定義し、この集合を特に意味クラスと呼ぶ。本研究では、節点のラベルは自立語に表すため、意味クラスは自立語集合として定義され

る。先述した例について考えると、共通変数に相当する節点のラベルは、次式のような意味クラス<色>を構成する。

$$\langle \text{色} \rangle = \{\text{黄色い}, \text{白い}, \dots\} \quad (1)$$

このように考えると、図1-aの定義文中において共通特徴および共通変数を表す部分は、図1-cのような意味クラスをラベルとする節点を含む部分グラフとして抽象化される。

ところで、部分グラフを、図1のように単一の節点に置換すると、これは、グラフ集合を対象とする一種の可逆圧縮となっている。ならば、辞書をグラフ集合と見なし、そのグラフ集合を効率よく圧縮することによって、辞書定義文からの特徴の抽出が可能であると期待される。

しかし、図1-b, 図1-cのように、部分グラフの選び方は何通りもの方法が可能なので、適切な選択基準が必要である。そこで、本論文ではMDL(Minimum Description Length)原理[1]を採用することにした。MDL原理における最良のモデルとは、モデルに基づいてデータを説明するためのパラメータの記述長と、そのモデルとパラメータを用いてデータを記述したときの記述長との和を最小とするようなモデルである。この考え方をグラフ集合の圧縮という問題に適用すると、部分グラフ集合を記述するための記述長と、その部分グラフを用いてグラフ集合全体を記述するための記述長の和を最小とするような部分グラフ集合を選ぶということになる。

3.3 辞書記述長の定式化

以下では辞書の記述長を定義し、「効率よく圧縮を行う」ということの定式化を行う。まずラベル付き有向グラフを一意に決定するための記述長を定式化する。次に、部分グラフの記述長、部分グラフに対応する節点を含むグラフの記述長について順に検討し、最後に辞書グラフ集合の記述長を与える。

グラフ集合の範囲 対象となるグラフ集合の範囲を確定するため、ラベル集合 Σ とタイプ集

合 Γ を定義する。まず、ラベル集合 Σ は、辞書に現れる全ての自立語と、意味クラスに対応する任意の自立語集合からなる集合である。例えば、図1-cに含まれる全てのグラフと部分グラフを記述できる最小限のラベル集合 Σ_{\min} は次の通りである。

$$\Sigma_{\min} = \{\text{花}, \text{さく}, \text{春}, \text{夏}, \text{白い}, \text{黄色い}, \langle \text{色} \rangle\}$$

また、タイプ集合 Γ は、辞書グラフ集合に現れる全ての辺のタイプと、辺クラスからなる集合である。ここで、辺クラスとは辺のタイプの任意の集合であり、意味クラスと同様に曖昧な部分グラフを表現するために導入した。集合 Γ の具体例は次の通り。

$$\Gamma = \left\{ \begin{array}{l} \phi(\text{連絡}), \phi(\text{無格}), \text{が}(\text{ガ格}), \dots \\ \langle \text{連絡} \rangle, \langle \text{無格} \rangle, \langle \text{ガ格} \rangle, \dots \\ \langle \text{連体修飾} \rangle, \langle \text{連用修飾} \rangle, \dots \end{array} \right\}$$

単純なグラフの記述長 ラベル集合 Σ とタイプ集合 Γ に属する要素によって表現される、部分グラフを含まない簡単な構文木 g を定義するために必要な記述長 $L(g|\Sigma, \Gamma)$ を定式化する。集合 Σ, Ω のもとでグラフ g が生起する確率を $P(g|\Sigma, \Gamma)$ とすると、適当な最適符号を適用することにより、次式が得られる。

$$L(g|\Sigma, \Gamma) = -\log P(g|\Sigma, \Gamma) \quad (2)$$

グラフ g に含まれる節点の集合を V_g 、辺の集合を E_g として、 V_g と E_g が独立であると仮定すると、次式が得られる。

$$P(g|\Sigma, \Gamma) = P(g|V_g, E_g)P(V_g|\Sigma)P(E_g|\Gamma) \quad (3)$$

ラベル集合 Σ が与えられた時、節点集合 V_g に属する全ての節点のラベルが一意にある状態に定まる確率を考えると、次のようになる。

$$P(V_g|\Sigma) = \prod_{v \in V_g} P(v|\Sigma) \quad (4)$$

$$\begin{aligned} P(v|\Sigma) &= P(v|C_v)P(C_v|\Sigma) \\ &= \frac{1}{|C_v|} \cdot \frac{1}{|\Sigma|} \end{aligned} \quad (5)$$

ただし、全てのラベルの出現が独立かつ一様であると仮定し、 C_v は節点 v のラベルである意

味クラスに含まれる自立語の集合とする¹。タイプ集合 Γ が与えられた時、辺集合 E_g に属する全ての辺のタイプが一意にある状態に定まる確率 $P(E_g|\Gamma)$ は、全ての辺の出現が独立かつ一様であると仮定すると、次のようになる。

$$P(E_g|\Gamma) = \prod_{e \in E_g} P(e|\Gamma) \quad (6)$$

$$\begin{aligned} P(e|\Gamma) &= P(e|C_e)P(C_e|\Gamma) \\ &= \frac{1}{|C_e|} \cdot \frac{1}{|\Gamma|} \end{aligned} \quad (7)$$

日本語文では、文末以外の全ての文節は、その文節よりも後方の1つの文節に対して係り受けを持つので、 $|E_g| = |V_g| - 1$ が成り立ち、可能な構文木の数は $(|V_g| - 1)!$ 通りとなる。グラフ g は、日本語の構文木と対応するグラフであるから、確率分布 $P(g|V_g, E_g)$ が一様分布であると仮定すると次式が成り立つ。

$$P(g|V_g, E_g) = \frac{1}{2^{|V_g|}} \cdot \frac{1}{(|V_g| - 1)!} \quad (8)$$

以上の議論より、グラフ g の記述長 $L(g|\Sigma, \Gamma)$ は次のように定式化される。

$$\begin{aligned} L(g|\Sigma, \Gamma) &= |V_g| \log |\Sigma| + \sum_{v \in V_g} \log |C_v| + \sum_{i=1}^{|V_g|} \log i \\ &\quad + |E_g| \log |\Gamma| + \sum_{e \in E_g} \log |C_e| \end{aligned} \quad (9)$$

部分グラフの記述長 部分グラフ s の記述長についても、同様に定式化される。ただし、図1-cの部分グラフ B のように、部分グラフ s の節点集合 V_s に意味クラスをラベルとする節点が含まれている場合、意味クラスに含まれるどの自立語が実際に現れるかを記述する必要はないため、式5に相当する式が次のように変更される。

$$P'(v|\Sigma) = \frac{1}{|\Sigma|} \quad (10)$$

部分グラフ s の辺集合 E_s に辺クラスをタイプとする辺が含まれている場合も同様に考える

¹ 節点 v が単なる自立語をラベルとする場合は $|C_v| = 1$

と、部分グラフ s の記述長 $L(s|\Sigma, \Gamma)$ は次のように定義される。

$$L(s|\Sigma, \Gamma) = |V_g| \log |\Sigma| + |E_g| \log |\Gamma| + \sum_{i=1}^{|V_g|} \log i \quad (11)$$

部分グラフを用いたグラフの記述長 次に、部分グラフ集合 Ω の要素をラベルとする節点を含むグラフ g の記述長を定式化する。この場合、部分グラフに対応する節点に係る辺についての構造的曖昧性が増加するため、式8に相当する式が次のように変更される。

$$P'(g|V_g) = \frac{1}{2^{|V_g|}} \cdot \frac{1}{(|V_g| - 1)! \cdot \prod_{v \in V_g} |V_v|^{E_v}} \quad (12)$$

ただし、 E_v は節点 v に係る辺の集合を表し、 V_v は節点 v と対応する部分グラフの節点集合を表す。なお、節点 v が部分グラフと対応しない場合、すなわち、節点 v のラベルが節点集合 Σ の要素であるような場合は、 $|V_v| = 1$ とする。よって、式2～7および式12から、グラフ g の記述長 $L(g|\Sigma + \Omega, \Gamma)$ として次式が得られる。

$$\begin{aligned} L(g|\Sigma + \Omega, \Gamma) &= |V_g| \log |\Sigma + \Omega| + \sum_{v \in V_g} \log |C_v| + \sum_{i=1}^{|V_g|} \log i \\ &\quad + |E_g| \log |\Gamma| + \sum_{e \in E_g} \log |C_e| \\ &\quad + \sum_{v \in V_g} |E_v| \log |V_v| + \sum_{v \in V_g} \log |E_v| \end{aligned} \quad (13)$$

ただし、節点 v が部分グラフと対応する場合は、 $|C_v| = \prod_{u \in V_v} |C_u|$ とする。

グラフ集合の記述長 したがって、辞書グラフ集合 D の記述長 $L(D)$ は次のように定義される。

$$\begin{aligned} L(D) &= \sum_{g \in D} L(g|\Sigma + \Omega, \Gamma) + \sum_{s \in \Omega} L(s|\Sigma, \Gamma) \\ &\quad + L(\Sigma) + L(\Gamma) \end{aligned} \quad (14)$$

特定の集合 Σ , Γ を選ぶ明確な基準は存在しないので、任意の Σ および Γ に対して記述長は等

しいと仮定すると、式14の第3項および第4項は無視することができる。その結果、

$$L'(D) = \sum_{g \in D} L(g|\Sigma + \Omega, \Gamma) + \sum_{s \in \Omega} L(s|\Sigma, \Gamma) \quad (15)$$

が目的関数となり、これを最小化するような集合 Σ , Γ と部分グラフ集合 Ω が、辞書の最も効率のよい圧縮であると考えられる。

4 アルゴリズム

この節では、3.3節で定式化した記述長を評価関数として、グラフ集合を実際に圧縮するためのアルゴリズムについて検討する。

MDL原理に基づく辞書グラフ集合の最適な圧縮は、式15の値を最小化するラベル集合 Σ , タイプ集合 Γ および部分グラフ集合 Ω を求めることと等しい。しかし、これらの集合の場合の数は非常に大きいため、全開探索は計算量的に不可能である。また、それぞれの集合の変化が相互に影響を与えるので、分割統治法や動的計画法などの効率的な探索手段を用いることもできない。

以上の理由から、本研究では記述長を最小とするような最適解を求めることを諦め、シソーラスから生成された静的な集合をラベル集合 Σ とし、文法に基づいてタイプ集合 Γ を決定する。その上で、出現頻度に基づくヒューリスティックスを用いた逐次改善法を適用して部分グラフ集合 Ω を探索し、近似的な解を求めることにする。

4.1 シソーラスに基づいたラベル集合

上述したように、最適なラベル集合 Σ を探索する問題は計算量的に困難であるから、人手によって構築されたシソーラスに基づく、固定されたラベル集合を利用することにした。

一般的に、シソーラスは木構造によって表現され、木構造の末端の葉は実際の単語を、中間の枝は意味分類を表す。これらの意味分類を、そのまま意味クラスと見なすことによって、静的なラベル集合を得る。例えば、図2のような簡単なシソーラスが与えられた場合、次のよう

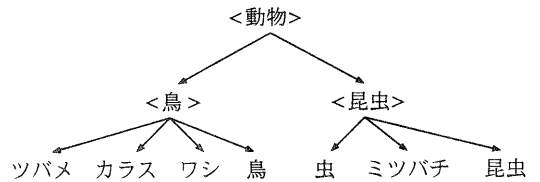


図2: シソーラスの構造

なラベル集合 Σ_1 が得られる。

$$\begin{aligned} \langle \text{鳥} \rangle &= \{ \text{ツバメ, カラス, ワシ, 鳥} \} \\ \langle \text{昆虫} \rangle &= \{ \text{虫, ミツバチ, 昆虫} \} \\ \langle \text{動物} \rangle &= \langle \text{鳥} \rangle \cup \langle \text{昆虫} \rangle \\ \Sigma_1 &= \left\{ \begin{array}{l} \text{ツバメ, カラス, ワシ, 鳥} \\ \text{虫, ミツバチ, 昆虫} \\ \langle \text{動物} \rangle, \langle \text{鳥} \rangle, \langle \text{昆虫} \rangle \end{array} \right\} \end{aligned}$$

4.2 ビーム探索

先述したように、最適な部分グラフ集合 Ω を求めることは計算量的に不可能であるため、逐次改善法を適用して、部分グラフ集合 Ω を探索する。すなわち、辞書グラフ集合 D から部分グラフを探索し、個々の部分グラフについて、それを部分グラフ集合 Ω に追加した場合の目的関数の値を計算し、最適値(最小値)を与える部分グラフを部分グラフ集合 Ω に追加する。目的関数の値が改善される限り、この操作を繰り返し、終了した時点の部分グラフ集合を解とする。

この時、部分グラフ集合 Ω に追加する候補として考慮されるべき部分グラフは、その時点のグラフ集合 D に含まれる任意の部分グラフである。しかし、対象となる部分グラフも非常に膨大なので、全ての部分グラフについて目的関数の値を計算することは不可能である。部分グラフの大きさに関するビーム探索と、出現頻度に基づくヒューリスティックスを組合わせた手順によって、部分グラフを探索した。

5 実験

上記の手法を用いて、実際にグラフ集合の圧縮を行った実験について述べる。

実験には、小学生用の日本語の辞書 [2] を用いた。これは、小学生用の辞書は、通常の大人用の辞書と比較して、限定された語彙のみを用いて丁寧に記述されていると考えられるので、定義文の構造に注目する本研究にはより適していることが期待されるからである。まず、簡単なフィルタを用いて定義文のみを抜き出し、JUMAN[3]による形態素解析とKNP[4]による構文解析を行い、個々の定義文をグラフに変換した。辞書の諸元は表1の通りであり、1つの定義文は平均して約3.4個の文節からなっている。ただし、計算機の記憶容量の制限から、辞書全体を対象とする実験は困難である。そのため、辞書の半分を削除し、シソーラスに存在しない単語を含む定義文を取り除くことにした。

表 1: 辞書の諸元

	全体	実験対象
見出し語数	28015	10087
定義文数	52919	13486
節点数	179597	53212

ラベル集合 Σ を生成するためのシソーラスとしては、分類語彙表 [5] を参照した。単語によっては複数の意味分類に属する場合があるが、その場合は登録されている複数の意味分類から1つを任意に選択し、それ以外は単に無視した。また、要素数が極端に大きい意味クラスが存在すると、出現数に基づくヒューリスティクスによる探索の障害となるため、シソーラスの中でも根に近い部分は削除した。次に、KNPの文法規則によって割り当てられた係り受けの素性を用いて、タイプ集合 Γ を決定した。

このようにして得られたラベル集合 Σ とタイプ集合 Γ を利用し、4.2節で説明したアルゴリズムを実行した。この結果、探索が停止するまでに、1409個の部分グラフが発見され、辞書全体の記述長は1153455.5bitから993121.6bitに低下した。これは13.9%の圧縮に相当する。得られた定義文の圧縮、すなわち定義文の特徴の例を図3に示す。

6 考察

図3の例に示すように、グラフ圧縮によって発見された辞書定義文の特徴は、ほとんどの場合我々の直感にあうものである。

「整とん」「整理」「身だしなみ」という3つの見出し語の定義文には、「きちんと<整理する>こと」という部分グラフが共通して現れている。「調髪」「乱心」などのように、「<整理する>こと」という表現の含まれている見出し語は他にも存在するが、その中でも特に適切な類似関係にある3語がこの部分グラフによってまとめられている。また、「きちんと」という表現が含まれている見出し語の集合（「几帳面」などの22単語）についても検討したが、妥当な単語がまとめられていることを確認した。

「いんげん豆」、「スパゲッティ」などからは「<調理して>食べる」という部分グラフがまとめられている。この部分グラフの含まれる見出し語は、いずれも食材であり、食材の調理方法に関する共通変数が適切なクラスによって取り出されている点に興味深い。また「食べる」という表現が共通している部分グラフとして、「<飯・そば>などにして食べる」という部分グラフがあり、これを含む見出し語もやはり食材である。「食べる」という表現が含まれている見出し語は他にも多数存在するが、食材に該当する見出し語は、この2つの部分グラフが含まれている8個でほとんど尽くされている。

「くちなし」「ほおの木」などの定義文には、「<色>花がさく」という部分グラフが共通して現れている。これらの見出し語は、<花の咲く植物>の要素として適切である。更に、<花の咲く植物>との上位下位の関係のみならず、色が異なるという点がきちんと取り出されていることが興味深い。ただし、この意味クラス<色>は形容詞からなるクラスであるため、「れんげ草」の定義文中に現れている「赤むらさき色の」のように「名詞+の」という組み合わせが形容詞的に働く場合は異なるものとして扱われてしまう。ラベル集合を生成する際に文法

調髪	髪を切りそろえて、形をととのえること。 ＜整備する＞	そうめん	ゆでて食べる。 ＜調理する＞
乱心	かなしみなどのために心が乱れること。 ＜整備する＞	たにし	煮たり、ゆでたりして食べる。 ＜調理する＞
整とん	きちんとかたづけること。 ＜整備する＞	ひじき	煮て食べる。 ＜調理する＞
整理	きちんとかたづけること。 ＜整備する＞	あさり	みそしるなどにして食べる。 ＜料理＞
身だしなみ	服装などをきちんと整えること。 ＜整備する＞	鰻	かばやきなどにして食べる。 ＜料理＞
几帳面	すみずみまできちんと正しいようす。	蛸	みそ汁などにして食べる。 ＜料理＞
いんげん豆	若いさやは煮て食べ、豆はあんこなどにする。 ＜調理する＞	くちなし	夏、かおりのよい白い花がさく。 ＜色＞
スパゲッティ	ゆでて、いろいろなソースであえたりして食べる。 ＜調理する＞	ほおの木	夏の初めに白い大きな花がさく。 ＜色＞
		れんげ草	春、赤むらさき色の花がさく。

図 3: 発見された特徴の例

的な知識を導入することによって、このような問題にも対処できる可能性がある。

7 結論

本稿では、辞書を知識源として利用する方法の第1段階として、意味の類似した一群の単語と、それらの相違点を発見する方法について述べた。最初に、実際の辞書の定義文についての考察から、辞書の定義文には、共通特徴、共通変数および個別特徴という3つの種類の特徴が含まれていることが分かった。次に、これらの特徴と定義文の構文木との関係について考察し、特徴の発見はグラフ圧縮の問題に置き換えて考えることができるという仮説をたて、特徴抽出を自動的に行うための方法を考案した。この方法は、MDL原理を基礎としている。実際に、日本語の小学生用の辞書を対象として実験を行い、この方法によって適当な特徴を自動的に発見することができることを示した。

今後は、得られた特徴をより詳細に検討すると共に、これを種々の自然言語処理アプリケーションにおいて利用することを検討する予定で

ある。

参考文献

- [1] J. Rissanen. *Stochastic Complexity in Stochastic Inquiry*. World Scientific Publishing Company, 1989.
- [2] 田近洵一（編）. 例解小学国語辞典. 三省堂, 1997.
- [3] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.6 使用説明書. 京都大学大学院 情報学研究科, 11 1998.
- [4] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6 使用説明書. 京都大学大学院 情報学研究科, 6 1998.
- [5] 国立国語研究所. 分類語彙表, 秀英出版, 1993.