構造化した属性表現に基づく国語辞典定義文の漸進的解釈

萩行 正嗣　　黒橋 禎夫
京都大学大学院 情報学研究科
{hangyo, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

国語辞典には語の様々な情報が整理された形で記述されている。国語辞典に記述された情報を獲得する研究は従来から行なわれてきたが、上位語、下位語、同義語などの定型的な表現によって記述されている情報の獲得が中心である。また、語には様々な多義性があり、獲得した情報を多義性を表現できる形で扱うことが重要である。そこで、本研究では獲得した情報をグラフ構造とすることで多義性を表現した。グラフ構造を利用したブートストラップ手法により国語辞典から様々な属性表現の獲得を行なった。

2　問題設定

国語辞典の定義文には見出し語の様々な情報が記述されている。以下の例では「焼き飯」の上位語が「料理」であり、材料が「飯」と「油」であることが分かる。

(1) 焼き飯：飯(材料) を 油(材料) でいためて作った 料理(上位語)。

本稿では〈焼き飯:材料:飯〉のような〈見出し語:属性名:属性値〉の3つ組の情報のことを**属性表現**と呼ぶ。ただし、以降では見出し語が自明の場合には見出し語を省略し〈材料:飯〉のように表現する。

本研究で扱う属性を表1に示す。上位語などは定型的な表現によりそのほとんどを獲得することができるが、多くの属性は定型的でない様々な表現で記述されている。例えば、材料であれば上記の例の「でいためて」の他に「を織って」「に加えた」など様々な表現がある。このような表現全てをルールとして人手で記述することは困難である。多様な表現に対する情報獲得の手法として Espresso[1] などの表層的な共起の情報を利用したブートストラップ手法が知られる。しかし、国語辞典ではある見出し語に対して高々数文しか定義文が存在しないため、従来のブートストラップ手法は利用できない。そこで、本研究では属性表現を利用したブートストラップ手法を提案する。まず、定型表現により記述されることが多い属性表現を獲得する(3節)。属性表現をブートストラップにおいて利用する際には多義性を考慮する必要があるため、獲得された属性表現をグラフ構造として扱うことで多義性を表現する(4節)。最後に属性表現を利用したブートストラップ手法を行ない属性表現の獲得を行う(5節)。

3　ルールによる属性表現獲得

国語辞典に記述された属性表現のうち一部は定型的な構文に対して制限を加えることで獲得できる。表2に獲得ルールの例を示す。このようなルールを人手で与えることで、国語辞典の定義文から属性表現の獲得を行なった。下記の例では「塁審」と「野球」に共通のカテゴリがなく、共に〈ドメイン:スポーツ〉をもつので、ルール3より〈塁審:サブドメイン:野球〉を獲得する。

(2) 塁審：野球で、一・二・三塁のそばにいる審判員。

4　属性表現の構造化手法

属性表現の扱いにおいて多義性による曖昧性は大きな問題となる。見出し語が多義である場合に獲得した属性表現がどの語義に対応する属性表現なのかが分からないためである。

語義曖昧性には大きく分けて2つの曖昧性があることが知られている。1つ目は対立的曖昧性と呼ばれるもので、意味的に関連のない語がたまたま同じ表記を持つために生じる曖昧性であり、日本語の場合には平仮名で書かれた語の曖昧性なども含む。例えば、バーには酒場という意味と棒という意味がある。

2つ目は相補的曖昧性と呼ばれるもので、意味関係を共有しながら文脈により異なる意味を示す語に生じる曖昧性である。例えば、「林檎」という単語は、果物の「林檎」を指す場合もあれば、果樹としての「林檎」を指す場合もある。これを属性表現で表現すると、前者は〈カテゴリ:食べ物〉、〈カテゴリ:植物〉、〈上位語:果物〉があり、後者には〈カテゴリ:植物〉、〈上位語:落葉高木〉がある。このように相補的な曖昧性がある単語に対しても、獲得された属性表現を一まとめに扱うのではなく、相補的な曖昧性を考慮して属性表現を扱う必要がある。

語義ごとに属性表現を分けて扱うために国語辞典の小見出しを一つの語義とすることが考えられる。しかし、国語辞典の見出しで語義を分割することは、対立的曖昧性と相補的曖昧性を同等に扱うこととなる。下記の「バー」の例では、小見出しで語義を分割すると、〈上位語:店〉と〈上位語:棒〉の区別と〈上位語:棒〉と〈上位語:横棒〉の区別を同等に扱うことになる。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 属性の種類

属性名	例 (見出し語:属性値)	定義
カテゴリ	うどん:食べ物、石:自然物	名詞の上位概念を 22 種類に分類 (JUMAN 基本語辞書に付与)
上位語	野手:人、塩:調味料	カテゴリの細分類となる表現
抽象物細分類	粒揃い:状態、カンパ:行為+成果物	カテゴリ-抽象物を 4 種類の大分類+2 種類の付加属性に細分類
移譲語	大統領:呼び声、小麦粉:原料	カテゴリの細分類としては不適な広義の上位語
ドメイン	野球:スポーツ、選挙:政治	名詞の属する話題を 12 種類に分類 (JUMAN 基本語辞書に付与)
サブドメイン	バッター:野球、税:税務	ドメインの細分類となる表現
目的	翼:飛ぶ、サイロ:貯蔵	その語の目的を示す表現
材料	風車:紙、素麺:小麦粉	その語の材料を示す表現
部分要素	蛸:吸盤、病棟:病室	その語の部分要素を示す表現
行為者	三振:打者、オリンピック:選手	その語の主体を示す表現
受け手	カルテ:患者、塾:子供	その語の客体を示す表現

表 2: 属性表現獲得ルールの例

	対象	制限	獲得する属性表現
ルール 1	定義文主辞	定義文主辞と見出し語に共通カテゴリがある または 見出し語のカテゴリが獲得されていない	⟨ 見出し語:上位語:定義文主辞 ⟩ ⟨ 見出し語:カテゴリ:上位語のカテゴリ ⟩
ルール 2	文頭の「A で、」	A と見出し語に共通カテゴリがある	⟨ 見出し語:上位語:A⟩
ルール 3	文頭の「A で、」	A と見出し語に共通カテゴリがない かつ (A と見出し語に共通するドメイン がある または 見出し語にドメインがない)	⟨ 見出し語:サブドメイン:A⟩ ⟨ 見出し語:ドメイン:A のドメイン ⟩
ルール 4	「A のため」	A が ⟨ 抽象物細分類:行為 ⟩ を持つ	⟨ 見出し語:目的:A⟩

(3) バー (小見出し 1)：棒。

(4) バー (小見出し 2)：走り高跳びや、棒高跳びで、跳びこす横棒。

(5) バー (小見出し 3)：酒を飲ませる店。

そこで、本研究では 2 つの多義性を表現するために、属性表現を構造化した**属性表現グラフ**を提案する。これは、属性表現をノードとし、関連が強いものは近い距離に、関連が弱いものは遠い距離 (または非連結) になるようにしたグラフ構造である。このような構造を構築するために国語辞典の小見出しの構造と属性表現獲得の際に制約等として利用した属性表現 (表 2 の下線部の属性表現など) の情報を利用する。

属性表現グラフのノードは属性表現 $a = \langle e : k : v \rangle$ とする。二つの属性表現 $a = \langle e : k : t \rangle$ と $a' = \langle e : k' : v' \rangle$ が以下の条件のどちらかに該当する場合に $edge(a, a') = 1$ とする。ここで $edge(x, x') = 1$ の場合グラフ上のノード x と x' がエッジで結ばれていることを示し、$x = x'$ の場合も $edge(x, x') = 1$ とする。

1. a と a' が同じ小見出しから獲得された
2. a を獲得する際に a' を利用した

このような条件で獲得した属性表現同士をエッジで結ぶことで属性表現グラフを構築する。

上述の「バー」では図 1 のような属性表現が得られる。この例では、「棒」という意味の属性表現と「酒場」という意味の属性表現とは非連結になることで対立的曖昧性を表現している。下記の林檎の例では二つの辞書からの定義文と ⟨ 林檎:カテゴリ:植物 ⟩ と ⟨ 林檎:カテゴリ:食べ物 ⟩ から図 2 のような属性表現が得られる。この例では、⟨ 林檎:上位語:果物 ⟩ を獲得する際に ⟨ カテゴリ:植物 ⟩ と ⟨ カテゴリ:食べ物 ⟩ が「林檎」と「果物」で共通なことにより、⟨ 林檎:上位語:果物 ⟩ が ⟨ 林檎:カテゴリ:植物 ⟩ と ⟨ 林檎:カテゴリ:食べ物 ⟩ に結ばれ、⟨ 林檎:カテゴリ:植物 ⟩ と ⟨ 林檎:カテゴリ:食べ物 ⟩ も結ばれる。このことから、食べ物であり植物であるという果物としての林檎の性質が表現されている。一方で、⟨ 林檎:上位語:落葉高木 ⟩ に注目すると隣接するノードは ⟨ 林檎:カテゴリ:植物 ⟩ だけとなっており、木としての林檎の性質が表現されている。

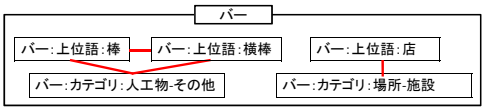

図 1: バーの属性表現グラフ

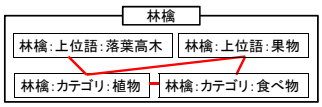

図 2: 林檎の属性表現グラフ

(6) 林檎 (rsk)：寒い地方で作られる 果物。

(7) 林檎 (iwanami)：ばら科の 落葉高木。

5 属性表現グラフを利用したブートストラップ型属性表現獲得

5.1 属性表現を利用したルール生成

国語辞典に記述されている属性表現のうち、材料、部分、行為者、受け手は多様な表現で記述されているため、人手で与えたルールによる獲得は困難である。そこで、本研究では獲得した属性表現グラフを利用したブートストラップにより、自動的にルールを生成することで国語辞典からの属性表現獲得を行なう。

国語辞典において 1 つの属性表現が定義文に記述されている回数は高々数回である。そのため、見出し

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

語と属性値の共起をインスタンスし、その出現した文脈をパターンとする従来のブートストラップ手法を用いることは困難である。そこで、属性表現を利用したルールを作成し、その定義文中での出現をインスタンスし、ルール同士の関連性をパターンとすることでブートストラップを行なう。ブートストラップによる獲得のシードとしては、定義文に対し人手によって属性表現獲得の正解を与えたものを利用する。

属性表現の獲得ルールとして 見出し語、獲得する候補の語 (候補語)、文脈毎にルール考え、その組合せを用いる。まず、見出し語と候補語のルールとしては、それぞれが属性表現として $\langle k:v \rangle$ を持つかどうかを、それぞれ $rule_e = \langle k:v \rangle$ と $rule_t = \langle k:v \rangle$ とする。文脈のルールとしては、候補語から連続して係る動詞に V があるかどうかをルールとし $rule_c = V$ とする。そして、これらの組合せたをルール $rule_n = (rule_e, rule_t, rule_c)$ とする。例えば、「焼き飯」(見出し語) の定義文から「飯」(候補語) にあてはまるルールは ($\langle$ 上位語 : 料理 $\rangle$, $\langle$ カテゴリ : 食べ物 $\rangle$, をいためて) や ($\langle$ カテゴリ : 食べ物 $\rangle$, $\langle$ カテゴリ : 食べ物 $\rangle$, をいためて) や ($\langle$ 上位語:料理 $\rangle$,$\langle$ カテゴリ:食べ物 $\rangle$, 作った) などである。

各ルールの属性表現 k らしさ $rule\_score(rule_n, k)$ をブートストラップにより計算する。

5.2 属性表現グラフを利用したルールグラフ生成

定義文中の候補語に当てはまるルールを属性表現グラフから網羅的に生成する。そのルールをノードとし、ルール間に関連性があればそれらをエッジで結ぶことでルールグラフを生成する。

まず以下のようなルールで見出し語 e、候補語 t、文脈 c の属性表現グラフを生成する。

見出し語 見出し語の属性表現グラフ

候補語 候補語の属性表現グラフ。ただし、表記により曖昧性がある場合にはそれらを合わせる

文脈 候補語から連続して係る動詞をエッジで結ぶ

下記の「瓦」の定義文で候補語を「ねんど」とする場合には、図 3 のようなグラフが生成される。

(8) 瓦：ねんど をかためて、かまで焼いて作る。

次に属性表現グラフのノードを 1 つずつ取り出して組み合わせることでルールグラフのノード $n = (a_e, a_t, a_c)$ を生成する。ノードを以下のように結ぶことで、関連するルール同士を結んだグラフを生成する。

$$edge(n, n') = 1 \; if \; (\forall x \in (e,t,c) \; edge(a_x, a'_x) = 1)$$

定義文グラフの例を図 4 に示す。各ノードの下の数字は説明のために便宜的に付与したノード番号である。ここで {1}〜{4} の「粘土」と {5}〜{8} の「年度」から生成されたノードは非連結となっている。このように、属性表現グラフを利用してルールグラフを生成することで、候補語にあてはまるルールをノードとし、関連するルールのノード同士を結ぶことができる。

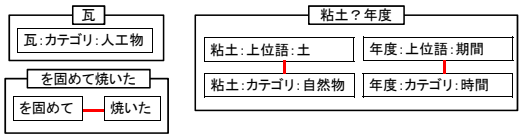

図 3: 属性表現グラフの例

図 4: ルールグラフの例

5.3 ルールグラフによる属性表現らしさスコア計算と属性表現獲得

ルールグラフのノードをインスタンスとし、隣接したノードをパターンとしてブートストラップ的に $rule\_score(rule_n, k)$ を計算する。まず、各ノード n の属性 k らしさを

$$score(n,k) = \underset{\{n' \in N | edge(n,n')=1\}}{\text{average}} (rule\_score(rule'_n, k))$$

によって計算する。ただし、人手によるシードとして与えられている定義文の場合には、⟨e,k,t⟩ が与えられた k, t の場合には全ての $score(n,k)$ を 1 に、それ以外の k, t については 0 とする。

全ての定義文においてルールグラフのノードのスコアの計算した後に、各ノードの $score$ から $rule\_score$ の計算を行なう。$rule\_score(rule_n, k)$ は以下の 2 つの式により計算する。

$$rule\_score_{exact}(rule_n, k) = \underset{\{n' \in N | a'_e \equiv rule_e \& a'_t \equiv rule_t \& a'_c \equiv rule_c\}}{\text{average}} socre(n'k) \quad (1)$$

$$rule\_score_{smoothing}(rule_n, k) = \sqrt[3]{\prod_{(x,y) \in \begin{pmatrix}(e,t),\\(e,c),\\(c,t)\end{pmatrix}} \underset{\left\{n' \in N | \begin{pmatrix} a'_x \equiv rule_x \\ \& a'_y \equiv rule_y \end{pmatrix}\right\}}{\text{average}} (score(n', k))} \quad (2)$$

ここで $a_x \equiv rule_x$ は属性表現 a_x が $rule_x$ を満たすことを表す。ブートストラップの初期 $rule\_socre$ には、

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

表 3: ルールベースによる属性表現獲得結果

属性	人手付与数	獲得数	適合率
カテゴリ	23174	20819	0.90
ドメイン	11377	111	0.85
上位語	198	30901	0.95
抽象物細分類	1116	29283	0.95
移譲語	10	9745	0.50
サブドメイン	99	1981	0.75
目的	0	2329	0.85

シードから式 (1) で計算したものを用い、それ以降用いる $rule\_socre$ は 6 節において実験により比較する。

上記の計算の繰り返しから最終的に計算された各ノードの $score(n,k)$ を利用して、候補語を属性表現として獲得するかを決定する。

$$(n_{max}, k_{max}) = \underset{n,k}{\operatorname{argmax}}\, score(n,k)$$

とし、$score(n_{max}, k_{max})$ が閾値を越えていた場合に、$\langle e : k_{max} : t \rangle$ を新たな属性表現として獲得する。ここで、各ノードが対応する候補語の語義の情報を持つので、属性表現獲得の際に曖昧性の解消が行なえる。例えば、「瓦」の例で $n_{max} = \{1\}$、$k_{max} =$ 材料 となり、$score$ が閾値を越えたとすると、$\{1\}$ の候補語の情報から、$\langle$ 瓦 : 材料 : 粘土 $\rangle$ を獲得することができ、平仮名表記の曖昧性が解消される。

6 実験

国語辞典として、岩波国語辞典と例解小学国語辞典を利用した。両方に含まれる名詞は 16565 語、岩波国語辞典のみに含まれる名詞は 30912 語、例解小学国語辞典のみに含まれる名詞は 3357 語であった。また、ブートストラップのシードとして Web 高頻度の語の定義文 2050 語に対して人手による属性表現獲得おこなったものを利用した。

まずルールベースによる獲得を 2 周行ない、その後にブートストラップによる獲得を 8 周行なった。ブートストラップによる獲得の際の閾値としては、各属性毎に人手で与えた属性表現が持つノード (5 節で $socre(n,k) = 1$ としたノード) が、式 (1) でスコアを計算した際に最も低いスコアとした。

6.1 実験結果

ルールベースにより属性表現を獲得した結果を表 3 に示す。ここで、適合率は無作為に選択した属性表現 20 個を人手で評価した結果である。

ブートストラップによる属性表現獲得の結果を表 4 に示す。exact は $rule\_socre$ を式 (1) で、smoothing は式 (2) で獲得した結果である。再現率は無作為に選択した 694 語に属性表現を人手で付与したもののうち実際に獲得できたものの割合である。

6.2 考察

ルールベースによる獲得では、属性表現による制限により、精度よく獲得することができたと考えられる。ブートストラップによる獲得では、材料を除き smoothing を行なった方が適合率、再現率共によい値となっている。材料においても図 5 から、同じ再現率で比較した場合には smoothing を行なった方が精度がよくなっている。これは、smoothing を行なわない場合には、シードとして与えた属性表現に過剰に影響されてしまうためである。また、スコアの閾値を下げることで適合率が下がることから、$rule\_score$ が属性表現らしさの指標として妥当であることが分かる。

表 4: ブートストラップによる属性表現獲得結果

属性	$rule\_score$	人手付与数	獲得数	適合率	再現率
材料	exact	274	915	0.822	0.235
	smoothing	274	1355	0.807	0.471
部分要素	exact	565	2021	0.440	0.190
	smoothing	565	2093	0.619	0.222
行為者	exact	181	446	0.689	0.181
	smoothing	181	605	0.700	0.181
受け手	exact	206	1560	0.463	0.217
	smoothing	206	1123	0.623	0.174

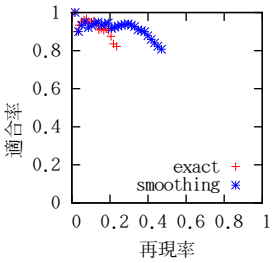

図 5: 材料の P-R 曲線

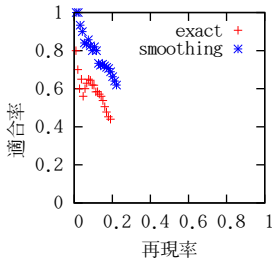

図 6: 部分要素の P-R 曲線

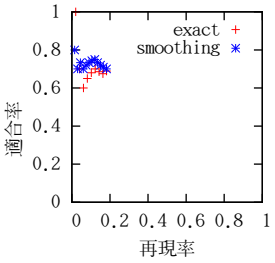

図 7: 行為者の P-R 曲線

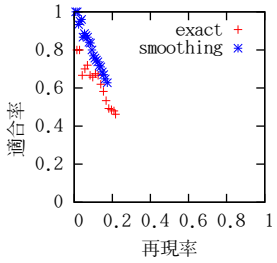

図 8: 受け手の P-R 曲線

7 おわりに

本論文では、属性表現のグラフ構造による提案し、属性表現グラフを利用したブートストラップ法を提案した。その結果、国語辞典という比較的小規模のコーパスにおいてもブートストラップ手法を適応することができた。今後の課題としては、属性表現の種類の拡張、形容詞などの修飾要素の属性表現グラフへの適用などを行いたい。

参考文献

[1] P. Pantel and M. Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc of COLING/ACL-06*, pp. 113–120, 2006.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.