EDR概念体系の抽象化と意味の弁別能力について

木村 和広　　平川 秀樹
東芝 研究開発センター
{kazu.kimura,hideki.hirakawa}@toshiba.co.jp

1 はじめに

意味的曖昧性解消 (Word Sense Disambiguation、以下 WSD と略す) 問題において、意味の単位をどの程度の細かさ (=粒度) にすべきか、またその粒度がある応用目的に照らして十分な記述力 (=意味的分解能) をもつのか、という問題を明らかにしていくことは重要な課題である。例えば、機械翻訳システム ALT[1] では約 3000 の意味カテゴリ (日本語のみ) が、WordNet[2] では、約 9 万の意味分類 (synset、英語のみ) が、EDR の概念体系 [3] に至っては約 40 万の意味分類 (日英共通) が用意されている。もちろん、粒度は詳細であればあるほど、その分解能は増すであろう。しかし、問題は、粒度が小さいほど、実際の意味的制約を開発するコストが級数的に増大することにある。さらには、人間の常識的な分解能を超えてしまうと、制約開発そのものが極めて困難な作業となってしまうこともある[1]。我々は、ある応用目的ごとに、それに応じた "適切な" 概念粒度が存在すると考えている。

本稿では、この問題について探るため、意味の上位下位関係を記述した概念体系に対し、一定の下位概念の集合を選びその集合を一つ以上の上位 (抽象) 概念に写像する手続きである、概念抽象化と呼ぶ操作を施すことにより、概念の抽象度 (=詳細性) と、概念間に成立する制約である概念記述の説明能力 (=妥当性・被覆性) との関係を実験的に明らかにする。概念体系や概念記述の初期値は EDR のものを用い、これに、抽象度をパラメータとして均等深度法・均等サイズ法・均等クラス確率法の 3 つの抽象化手法を適用する。そして、EDR コーパスから抽出した共起表現に対し、抽象化された概念体系と概念記述を用いて、WSD の実験を行い、抽象度の高さと概念記述の妥当性・被覆性の関係を求めると共に、抽象化手法間の比較を行う。

[1] 我々は、EDR コーパスの意味タグの品質が十分と言えない原因は、ここあると考えている。

2 概念抽象化

2.1 概念抽象化の意義

概念抽象化は、概念の上位下位関係を記述した概念体系をベースとし、機械的操作により、下位概念を上位概念に写像することで、体系を簡略化するものである。例えば、概念< サラブレッド >を概念< 馬 >や概念< 動物 >に抽象化することで、体系化される概念ノード数は減少し、体系保守の負荷が軽減される。さらに、適切な抽象化は、無用な細分を抽象化してしまうことで、概念記述の開発に適したきめ細かさ (=概念粒度) を提供し、概念粒度の均一化が期待できる。また、個別事例から得られた概念記述は適切なレベルに一般化され、他の事例に適応し得る知識となる。例えば、概念記述< サラブレッド > − < *agent* > − < 走る > は、< 馬 > − < *agent* > − < 走る >のように抽象化され、"サラブレッド" 以外の馬にも適用しうる知識として一般化できる。一方で、過剰な抽象化は、誤った知識を導くことに留意する必要がある。

2.2 概念抽象化の手法

本稿で取り上げる概念抽象化の手法は、以下の 3 手法である。

均等深度法　ルート概念からの深さが定数 D 以上である概念を深さ D の上位概念に抽象化する。

均等サイズ法　ある概念が支配する概念数(子孫となる概念数)が一定値Sとなるように子孫概念を祖先の概念に抽象化する。

均等クラス確率法　ある概念の出現確率を考えたとき(上位概念の出現確率は自身の出現確率に支配概念の出現確率の総和を加えたものとする)、その出現確率が一定値Pとなるように子孫概念を祖先の概念に抽象化する。

2.3　概念抽象化における多重継承の問題

概念抽象化を考える上では、多重継承(=ある概念に対し複数の上位概念の存在を許し、それら上位概念の属性をすべて継承する)を受けた概念をどのように抽象化すべきか、という問題が発生する。

本稿では、束構造を木構造に展開して考え、同一概念をルート概念からのパスによって別概念と見なす、という考え方をとり、複数の上位概念すべてに抽象化を行うことにした。この場合、1対多の抽象化を行うため、概念記述の抽象化において無用な曖昧性を生ずる(本来どれか一つに抽象化されるべき概念記述をすべてに抽象化してしまう)という問題があるが、概念記述に重み(頻度)を与えることで、誤った抽象化の重みが小さくなることが期待される。

3　EDR概念体系を用いた概念抽象化実験

本節では、EDR概念体系[3]を用いた概念抽象化実験について述べる。

EDR概念体系は、他のEDR辞書群(単語辞書、対訳辞書、共起辞書・コーパス)と概念識別子と呼ばれる16進数を媒介として、相互に関連付けられている。単語辞書は、単語(表層)情報と概念情報との対応を与える。EDR概念体系は言語依存性(日本語・英語の別)をもっていない。

概念抽象化にあたっては、まず、体系の日本語化を図る必要があった。これは、均等サイズ法や均等クラス確率法においては、着目しない言語(今回は英語)に関するノードの影響が問題となるからである。体系の日本語化により、英語単語辞書のみに対応づけられた葉ノードおよびそれらのみを支配していた中間ノードがすべて削除され、ノード数199245個の体系となった[2]。

次に日本語化概念体系に対し、均等深度法を適用する。すなわち、体系をトラバース(depth first)しながら、深さが指定値を下回るノード n を発見した時点で、ノード n 以下全支配ノードを n の直上ノードに抽象化する。抽象化の結果は、元の概念識別子と抽象化後の概念識別子(複数可)との対応表として保存する。この対応表を以後**抽象化マップ**と称する。均等深度法では、最深深度の概念ノードは深さ16に体系づけられているため、depth=1から15までの抽象化マップを作成した。

均等サイズ法では、体系上の全ノードに対し、支配ノード数を算出(前節で述べた木構造展開の考え方から、多重継承はダブルカウントする)してDB化する。次に、体系をトラバースしながら、支配ノード数が指定値を下回るノード n を発見した時点で、ノード n 以下全支配ノードを n の直上ノードに抽象化。但し、抽象化ノード間にisa関係が成立する場合は、より子孫である方のノードのみ採用する(これを最小抽象化と呼ぶことにする)。均等サイズ法では、size=1から50000までの15段階について、抽象化マップを作成した。

均等クラス確率法では、まず、EDRコーパスに出現した自立概念(自立語の概念)の頻度をカウントする。なお、sparseness対処として、Good–Turing discount法[4]により、観察頻度を一定量間引き、間引いた頻度を観察されなかった事象(非出現概念)に均等に分配する。次に、体系上の全ノードに対し、頻度を算出(支配ノードの頻度を合算、多継承は親の数で均等に分配)してDB化する。そして、体系をトラバース(depth first)しながら、上記頻度が指定値を下回るノード n を発見した時点で、ノード n 以下全支配ノードを n の直上ノードに抽象化し、さらに最小抽象化を行う。均等クラス確率法では、probability=0.00001から0.2までの17段階について、抽象化マップを作成した。

[2] EDR概念体系V1.5では、未分類概念(体系化の行われていない雑多な概念)が、体系上の上位レベルに配置されており合わせてこれらも削除した。

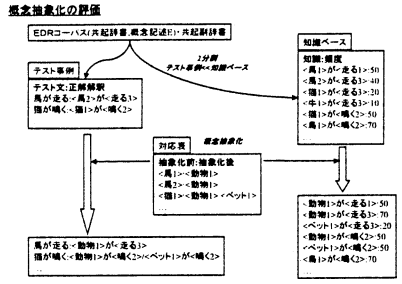

図 1: WSD タスクによる概念抽象化の評価

4 意味的曖昧性解消タスクによる評価

4.1 評価方法

概念抽象化の評価として、意味的曖昧性の解消 (WSD) タスクを設定する。

WSD タスクの概要を、図 1 に示す。本タスクの基礎データには、EDR コーパスから共起情報のみを抽出・整形したデータである EDR 共起辞書を用いる。EDR 共起辞書は、共起表現、すなわち、2 つの共起する単語 (表層語) とそれらを結ぶ共起関係子 (格助詞等) を見出しとして整列させたものであり、特にその共起表現の意味的解釈であるところの概念記述、すなわち、その 2 単語の概念とその 2 概念を結ぶ概念関係子が記述されている。すなわち、表層表現に対し意味的曖昧性を解消した結果 (=正解データ) が多数収録されたものである。

本評価では、この共起データをテスト事例と知識ベース用事例の 2 つの部分に分割する。テスト事例は、WSD テストのテストデータであり、知識ベース用事例は、WSD のための知識源である。知識ベース用事例からは、正解解釈である概念記述のみを抽出し、それを出現頻度 (これを概念記述の信頼度=スコアと仮定する) つきで DB 化することで、知識ベースを構成した。なお、スコアは、頻度と体系上の位置を考慮し、次式で与えた。

$$score(< rel, c1, c2 >) = \frac{freq(< rel, c1, c2 >)}{dom(c1) \cdot dom(c2)} \quad (1)$$

ここで、$freq(rel, c1, c2)$ は概念記述 $< rel, c1, c2 >$ の出現頻度、$dom(c)$ は、概念 c の支配ノード数である。これは、概念記述の出現頻度を当該概念の支配ノード数で割ることで、葉ノード間の関係として概念記述を分割し、その頻度を均等に分配したのと等価な効果が現れるように配慮したものである

今回は、テスト事例として約 1.4 万件[3]を選び、これを約 50 万の知識ベース用事例により WSD 実験を行った。

WSD テストの概要は以下の通りである。これを抽象化を行わない場合 (=ベースライン) と、前節に述べた抽象化条件 47 通りに対し行う。

step 1 前節で作成した抽象化マップを用い、知識ベースを抽象化する。抽象化が n 通り存在した場合は、元のスコア s に対し、抽象化概念記述のスコアには s/n を与える。

step 2 各テスト事例に対し、可能な解釈を洗い上げ、概念階層の継承関係と上記抽象化知識ベースを用いて、成立する解釈をスコア付で選択する。スコアは、照合した抽象化知識のうち最大のものを採用する。成立する解釈が一つでもある場合、WSD が行われたと呼び、カウンタ c でこの数を数える。

step 3 上記の可能な解釈のうち、スコアがトップである解釈 N 個 (スコアが同点のものが N 個あればすべて候補とする) の中に正解解釈があれば、$1/N$ をそのテストデータに対する正解率 p_j とする。

step 4 下記の式により、正解率 P、被覆率 C、F 尺度 F を求める。ただし、N はテスト数である。また、F 尺度における定数 β は適合率に対する再現率の相対的重要度 (β=1 は両者の重要度が同等) であり、今回の評価では、β=1 で算出した。

$$P = \frac{\sum_j p_j}{c} \quad (2)$$

$$C = \frac{c}{N} \quad (3)$$

$$F = \frac{(\beta^2+1)CP}{\beta^2 P + C} \quad (4)$$

[3]関係子は agent,object のみ、曖昧性平均 16.4

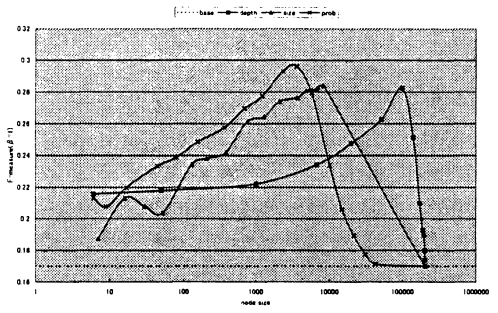

図 2: 概念抽象化による WSD 能力 (F 尺度)

4.2 評価結果

評価実験の結果を図 2 に示す[4]。各抽象化手法の違いを見るため、横軸には抽象化されたときの概念ノード数をとっている。図中ベースライン評価の結果を点線で示している。

正解率・被覆率について ベースライン正解率は、33.6%であった。もっと高い正解率が予想されたが、知識ベースのスパースネスが原因と考えられる。手法同士の比較では、均等クラス確率法、均等サイズ法、均等深度法の順に少ないノード数でベースラインに収束した。つまり、均等クラス確率法は最も少ないノード数でベースライン並の意味の弁別能力を備えた概念体系を与えたことになる。一方、ベースライン被覆率は、11.4%であった。抽象化が高いほど被覆率は高まり、正解率とは逆に、均等深度法、均等サイズ法、均等クラス確率法の順に少ないノード数で 100%に収束した。

F 尺度について ベースライン F 尺度は、17.0%であった。各抽象化手法は、被覆率は上げるが正解率は下げる方向に働くのが基本であるが、総合的にみれば、どれもベースラインを上回る結果を得た[5]。手法同士の比較では、極大値が、均等クラス確率法・均等サイズ法・均等深度法の順に少ないノード数で現われており、均等クラス確率法がベストの結果といえる。均等クラス確率法の極大値は 29.6%で、このときのノード数 3583 であった。これまで、経験的指標として、体系を構成するノードは 3000 程度が適当であるとの見解があった。この極大値を与えるノード数は、この数値に合致する。

[4] 紙面の都合上、正解率・被覆率の結果は割愛する。
[5] もちろん β の取り方に依存する。

5 まとめ

本稿では、均等深度法・均等サイズ法・均等クラス確率法の 3 つの抽象化手法を用いた概念抽象化を行い、WSD タスクによる評価を行った。その結果、どの抽象化手法においても、抽象化を行わない場合と比べ、総合的にみて優位な意味処理能力を有することが明らかになった。また、手法同士の比較では、概念の出現頻度を基礎として行う均等クラス確率法がベストな結果を生んだ。

今後、目視評価を加えて、知見を深めてゆくとともに、評価により明らかになった概念記述のスパースネスを埋めるための方策を検討してゆく予定である。

参考文献

[1] 池原悟, 宮崎正弘, 横尾昭男: 日英機械翻訳のための意味解析用の知識とその分解能, 情報処理学会論文誌, Vol.34, No.8, pp.1692–1704 (1993).

[2] Christiane Fellbaum(Eds.): WORDNET – An Electronic Lexical Database, The MIT Press (1998).

[3] 日本電子化辞書研究所編: EDR 電子化辞書仕様説明書 (第 2 版), EDR Technical Report TR-045 (1995).

[4] Good,I.J.: The population frequencies of species and the estimation of population parameters, Biometrika, 40, pp.237–264 (1953).

[5] 内山将夫, 板橋秀一: シソーラス上に動的に構成される標本空間における動詞の多義性解消, 自然言語処理, Vol.4, No.3, pp.27–50 (1997).