

EDR 電子化辞書における概念体系

荻野孝野 仲尾由雄 長澤陽子 小笠原あゆみ
日本電子化辞書研究所

1 はじめに

本報告は、日本電子化辞書研究所(略称 EDR)における電子化辞書プロジェクトの成果のうち、概念体系部分の開発の詳細を報告するものである。EDR には、基本語辞書の日英各 20 万語に配置された約 40 万概念のセットがある。我々は、この 40 万の概念を対象を、体系化された意味分野の枠組みを作成し、それぞれの概念を、この枠の中の適切な場所に分類した。

この体系は、約 6000 の分類枠からなる。分類は、単語ではなく、概念を対象として分類を行ったものであり、この点において、従来の、単語を対象とし、ありうる概念を想定しながら分類を行う作業とは大きく違うところである。ただし、分類枠の作成においても、概念の分類においても、様々な問題が生じた。ここでは、これらの問題についても報告する。

2 EDR 概念体系の概要

2.1 体系作成の基本方針

何らかの観点において、意味的な関連を持つ単語あるいは概念をグループ化し、グループ化された集団どうしを意味的な関連から構造化して配置したものが概念体系である。意味的な関連を持つ単語をグループ化するにも、さまざまな観点がある

1) 言語外関係によるグループ — 指示物自体が持つ知識的な体系

人事における階級とか、動物の生物学的分類とか、言語内部の表現とは無関係に指示物自体が、何らかの知識体系を持っている場合である。

例 1 {鉛筆, ノート, 消しゴム, カッター} < [文房具]

例 1 にあげた単語は、知識としては、[文房具] の範疇に入るものである。我々は、これを、分野的分類と位置づける。[家庭用品]、[農具] などと言った束ね方が、この分野的分類に該当する。

2) 言語内関係によるグループ — 表現上の意味的関連でとらえた体系

文内の他の語との関連において、似かよった働きをするものから意味の共通性や差異を明らかにし、表現上の共通性によってグループ化するものである。

例 2 {小刀, 刀, 包丁, カッター, はさみ} < [刃物]

例 2 にあげた {カッター} は、1) の分類によると [文房具] のグループであるが、2) の分類によると、[刃物] のグループである。

[刃物] に入る語群は、{切る, 切れる, 鋭い} などと密接な関係を持つが、[文房具] に入る語群は、使ったり買ったりする場所に関連が出てくるくらいで、述語側に強い関連を持つ特定の語群を想定させるものではない。

このように、分類の観点の違いによって、全く違った単語グループができる。EDR では、概念体系作成の主たる目的を、概念記述の補完や、文表現上の機能から見た意味的類似度の推定などにおいたため、分類は、2) の言語内関係の観点を主とし、まとめる概念(上位概念)が単語にリンクしている実概念として存在するようなものについては、1) の知識的分類の観点も含めた。

2.2 形式

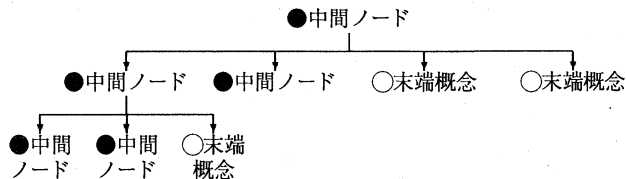
EDR 概念体系は、概念と概念を上下関係で関係づけたもので、以下のような形で表現されている。

下位概念 —kind-of→ 上位概念

同じ上位概念に属する下位概念どうしが、親の意味成分を共有する、同じ親の下の子供どうしということになる。

- 1) 体系の深さ: 基本的に深さの限定はない。
- 2) 下位概念の数: 体系の中間ノードの下位に属する下位概念の数の限定はない。
- 3) 下位概念のレベル: 下位概念には、さらに下位概念を持つ中間ノードが続く場合と、それ以上、下位概念を持たない末端概念が続く場合とのいずれも可能である。

図1 体系上における中間ノードと末端概念



各レベルにおける末端概念は、そのレベルに用意された中間ノードに入らない概念で、特殊で雑多な概念であることが多い。

2.3 数量的概要

表1 EDR 概念体系の数量的概要

中間ノード数	6,221
分類の対象となった概念数	400,000
体系の深さ(最大)	14

3 EDR 概念体系の特徴

3.1 体系の組み立て

はじめに述べたように、何らかの共通な意味特徴を持っている単語を集めるのが概念分類、それらのグループ間の意味的差異に着目して、階層的に配置するのが概念体系である。

体系の組み立てには、概念の包摂関係が深くかかわる。包摂関係は、異なる概念の間に共通する意味成分の組み合わせによって、類似性が見られるものを束ねる。さらに、意味成分の組み合わせを少なくすることによって、意味的に含む範囲を緩やかにし、概念上の上位関係を作っていくものである。

図2. 意味成分の組み合わせと上下関係

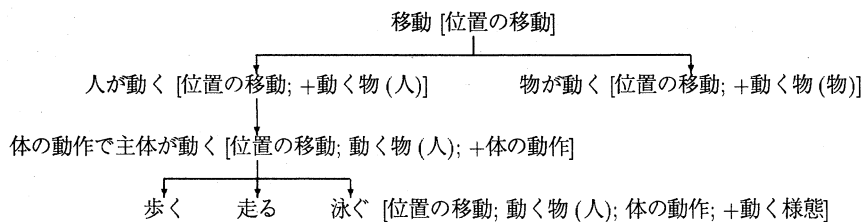


図2で、上下方向にたどって「+」部分が意味成分の追加で、上位概念をより限定するものであり、同じ階層の横方向部分の「+」部分の値の差が下位概念どうしの概念の差になる。

3.2 体系の構造

EDRの概念体系は、概念記述の支援としての役割を想定したため、基本枠組みは「文中における意味的役割が同じ振舞いとなる概念」をグループ化することにより組み立てた。

文における構文的役割を担う意味のまとまりを体系の上位部分とし、表現の役割に従って、

1) [主体] (働きかけの主体となりうる意志を持っているものとみなされるもの)

2) [働きかけあるいは変化の対象]

3) [主体あるいは対象の動きや状態]

および [主体あるいは対象の動きや状態を限定する時空間] の意味領域に分けた。

次に、[主体] については、「実体があるものかどうか」(具体的有意志体 {人間, 動物, 計算機や自動車などの自動物}; 抽象的有意志体 {組織, 会合, 神ナド}), 「実体があるもの」については、「生きているかどうか」(生物 {人間, 動物}), 「自ら動くことができるものかどうか」(可動 {人間, 動物, 自動車}), といった観点の組み合わせで、下位概念をまとめている。これらは、

[主体] (属性; 実体, 可動) — {動く}

[主体] (属性; 実体, 可動, 動物) — {歩く}

[主体] (属性; 有意志体) — {考える}

というように、述語との関係において、有効な意味成分の組み合わせを抽出しやすいような中間ノードを作成したものである。

図3. 概念体系の概要

- 1) [人間または人間に準ずる主体]
 - [人間]
 - [動物]
 - [自立活動体]
 - [仏, 霊, 天使など, 抽象的な有意志体]
 - [市場, 世間, 社会など, 人間にかかわる観念的な領域]
- 2) [活動の対象や道具や作り上げたもの]
 - [もの]
 - [具体物]
 - [抽象物]
 - [具体物, 抽象物のいずれでもありうるもの]
 - [自然現象物]
 - [事柄]
 - [遊びやスポーツ]
 - [催し]
 - [出来事]
 - [社会, 文化, 政治などにかかわる運動や改革]
 - [政治, 経済]
- 3) [事象]
 - [現象]
 - [行為]
 - [移動]
 - [変化]
 - [状態]

4 作成過程における諸問題とその扱い

4.1 概念記述向き分類と類語的分類のずれ

例3 C#嫁ぐ {行く}

これは、類語的分類であれば [結婚する] のグループに入り、概念記述向き分類であれば [終点への移動] に入る。前者は [人どうしの相互行為] で、同等の [主体 (AGENT)] を二つ要求するパターン。後者は [主体が終点へ移動する] で、[主体の移動] がかつ [終点 (GOAL)] を要求するパターンに属する。このようなずれが生じた場合には、基本的に概念記述向き分類に従った。

4.2 多重分類

4.2.1 分類軸の多様性による多重分類

一つ概念を下位概念に分類する際、意味特徴を抽出するための「観点」=「分類軸」の設定のしかたで複数の分類が可能になる場合がある。例えば、[具体物] を分類するとき、[生物/無生物] [自然物/生産物] [材料/道具/加工物] などは、それぞれが分類軸として採用可能であり、多重分類が生ずることになる。

例4 C#野菜 < [食品] (~を食べる)
< [植物] (~が育つ)

4.2.2 意味役割の多様性による多重分類

本来ならば、単語の多義性はすべて列記されることが望ましいが、同じような意味機能を持つ場合は、一般の辞書でも多義としては区別されていない。

EDRの辞書では、{学校}については、1) 学校という教育を行う組織、2) 学校という教育施設、という二つの語義が立てられているが、{小学校、大学、専門学校、...}などについては、「学校という教育施設」しか記載されていない。そこで、同じような意味特徴を持ち、文中において異なる意味機能として働く場合は、体系上の多重とした。

例5 C#学校 < [建物] < [生産物]
..... < [場所]
< [組織]

4.2.3 属性や評価による修飾部分を含むものの多重分類

例6.1 C#美しい人 {美人} < [外見で捉えた人間] < [パーソナリティで捉えた人間]
< [人の外見的印象の値]
例6.2 C#珍しいもの {珍物} < [価値の評価で捉えた具体物] < [評価で捉えた具体物]
< [価値の値]

例6.1と6.2に示したように、属性や評価による修飾部分を含んだ概念は、構文上主語であれば「C#もの」として、述語であれば「C#もの」の属性や評価を示すものとして用いられる。このように、「形状」、「価値」など、属性や評価で限定された「もの」は、構文上の位置によって、担う意味的役割が異なる。このタイプについては、4.2.2と同様、複合的な概念で束ね、概念体系上で展開する形を想定し、複合的な概念を表す分類名にした。

4.3 概念表現の曖昧性と分類

EDRの概念分類は、単語ではなく、概念見出しを解釈しながら行ったものである。この方法によると、概念分類の精度の善し悪しは、概念見出しの書き方が的確であるかどうかによって依存するところも大きい。概念分類するにあたって、概念見出しと単語の示唆するもの間にずれがあったものなど、分類上、問題が生じた部分もある。

例7 C#人に酒をすすめる {差す}

概念見出しだけに依存して分類すると、[勧める] (< [情報の移動を伴う対人行為])、単語によって分類すると、[液体を注ぎ入れる] (< [具体物を始点から終点へ移動する行為]) となる。ここでは、[注ぐ]の方に重きをおいて分類した。

5 おわりに

大量の概念分類のため、分類作業も多年にわたり、部分的に不十分な点もあると思うが、本概念体系が今後の自然言語処理における概念関係データの叩き台たりうれば幸いである。

文献

- [1] 荻野孝野, 仲尾由雄, 小笠原あゆみ, 長澤陽子, 「日本電子化辞書研究所における概念体系」, (情報処理学会研究報告 Vol.93, N0.98), 1993.
- [2] 日本電子化辞書研究所 TR-041, 「EDR 電子化辞書仕様説明書」, 1993.
- [3] Eugene A. Nida 著, Noah S. Brannen 監訳, 升川潔, 沢登春仁訳, 「意味の構造 - 成分分析」, 研究社出版, 1977.