EDR 日本語辞書からの情報獲得のための概念説明文の解析

藤原滋

北陸先端科学技術大学院大学 情報科学研究科

1999 年 2 月 15 日

キーワード： EDR 電子化辞書, EDR 概念体系, 概念説明, N-gram 頻度統計.

近年，大規模な機械可読辞書やコーパスによる自然言語処理が行われている．EDR 電子化辞書は (株) 日本電子化辞書研究所の製品で，通常いわれる辞書に加え，シソーラス，コーパスを含む，機械可読な言語データベースである．EDR 電子化辞書の一部である EDR 概念体系は，40 万概念を上位下位関係によってまとめた大規模なシソーラスである．EDR 概念体系からは，上位概念，下位概念，類義の関係にある概念を得ることができる．これらの情報を必要とする処理にとって，EDR 概念体系の持つ豊富な概念は有用である．

ただし EDR 概念体系を利用する処理にとって，概念体系の性能は必ずしも十分ではない場合がある．例えば，

1. 概念体系中の概念の分布に偏りがある．ある分野についてより豊富な概念を利用したい．

2. EDR 概念体系から得られる兄弟概念は多すぎて，その処理にとって必要のない概念まで得られてしまう．EDR 概念体系の与える兄弟概念よりもより細かい類義の概念がほしい．

という要求がある．このうち 1は，辞書編纂者による概念の追加により解決するが，2は，要求水準が一般的には利用する処理毎に異なることから，全ての要求を満たすようなシソーラスの構築は現実的ではない．この要求を満足させるために，**各概念に兄弟概念間の差異を明確にするような付加情報を与える**という方法が考えられる．

EDR の概念には既にそのための情報として**概念説明**がふられている．概念説明は，概念の意味内容を自然言語による一文で説明したもので，通常の辞書でいうところの語釈文に相当する情報である．しかし，概念説明は自然言語で記述されているため，そのままで

は概念説明のもつ情報を計算機から十分に利用することは難しい．そこで本研究は EDR 日本語概念説明を対象に，概念体系から直接得られる兄弟概念より詳細な類似概念を得るために概念説明からどのような情報が得られるかについて検討し，それらの情報を抽出するプログラムを実装し，評価を行った．

まず本研究では概念説明と辞書の語釈文との類似性に着目し，概念説明の表層の特徴から文の構造についての情報を得ることを試みた．概念説明を，その概念を語義としてとり得る語の品詞で分類し，各々の概念説明の集合に対し N-gram 頻度統計をとった．その結果，品詞毎に異なる特徴を持った頻出表現を得ることができた．さらにいくつかの頻出表現は概念説明特有の文の構造を解析する際に強力な手がかりを与えることがわかった．本研究ではそのような頻出表現 (手がかり語) の情報から，概念説明で意味的に重要な役割を担っている語として**定義語**を，特徴的に使われ，意味的な役割がはっきりしている文のブロックとして**「という」部**，**「において」部**を提案した．定義語,「という」部,「において」部は，手がかり語の情報と汎用の形態素解析器，構文解析器を用いて容易に抽出することができる．

本研究で提案する概念説明の意味解析の枠組を説明する．まず，親概念を同じくするグループを作り，これを処理の最大の単位とする．次に兄弟概念のグループ内で，同じ語をまとめたグループと，同じ受け語に同じ格で係っている語のグループを全てまとめあげる．以上のようなグルーピング処理をあらかじめ行った上で，定義語,「という」部の語,「において」部の語に対して，**N-gram 頻度統計から得られた手がかり語の情報による意味的制約の利用**，**EDR 概念体系の情報を利用したスコアリング**により語義を決定する．同じ語のグループ内の語で既に語義の決定されたものがあれば，グループ全体の語義を fix する．同じ受け語に同じ格で係っているような語のグループに対しては，**それらが似たような意味の語である**という仮定に基づく手法で語義を決定する．ここまでで語義の決まっていない語を含む全ての係り受けの組に対して，EDR 共起辞書の頻度情報に基づくデフォルトの語義決定手法を適用する．

本研究で提案した手法の有効性を評価するために，実験を行った．EDR 概念体系の葉ノードである概念の一段上位の概念から 21 概念を無作為抽出し，それぞれに対して，その概念を親概念とするような兄弟概念で，且つ日本語概念説明を持つ概念 (計 535 概念) の集合を評価セットとした．その結果，定義語,「という」部,「において」部の自立語に対しては，本研究で提案した手法により，recall 値 56%，precision 値 64% という比較的高い精度で語義の決定を行うことができた．手がかり語の情報が利用できない場合のデフォルトの語義決定手法の精度は recall 値 33%，precision 値 37% であった．そして，これらの語を含む評価セット全体の語義決定の精度は recall 値 60%，precision 値 55% であった．

特に定義語,「という」部の語,「において」部の語に対して本研究で提案した手法はコストの低い手法の組合せにも関わらず，比較的高い精度で語義を決定できることがわかった．

今回全体的に精度はそれほど高くはなかったが，スコア自体にまだ改良の余地があるほか，2 位以下のスコアの語義も語義として採用するなどの工夫によってもまだ精度を向上する余地があるものと思われる．