修 士 論 文

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

指導教官　奥村学 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

藤原滋

1999 年 2 月 15 日

要 旨

EDR 電子化辞書は 40 万概念からなる概念体系 (シソーラス) をもち，それらの概念を説明するものとして，各概念に対して概念説明文がふられている．しかし，概念説明は自然言語で記述されているために，現在は概念説明文が持つ情報を計算機から直接には利用できない．

本研究では，自然言語で記述されている EDR 日本語概念説明文の持つ情報を計算機から利用できるようにするために，形態素解析，構文解析，意味解析を行う．最終的には EDR 日本語概念説明文中の形態素区切りの 情報，形態素間のかかり受け情報，文中の自立語がどのような語義で出現しているかという意味情報を得るのが目標である．

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 研究の背景

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 [2] 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 (図 1.1) [1]．概念自体がどのようなものであるかを説明するものとして，概念見出しと，**概念説明**がある．概念見出しは概念の意味内容を一単語で説明したもの[2]，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ある[3]．概念説明は，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．概念見出し，概念説明は，全ての概念に付与されているとは限らない[4]．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 (以降，親概念，子概念，兄弟概念と呼ぶことがある) 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．

[1] EDR の概念自体は直接語と結び付いているものではないが，EDR 単語辞書が語と概念間のリンク情報を提供する．

[2] 互いにごく近い概念の間では，概念は異なっていても，概念見出しは同じである場合がある．すなわち概念見出しは一意性が保証されない．

[3] 概念説明は一意である．

[4] 高度に抽象的な概念では見出しまたは説明がない場合がある．また，英語の概念見出しや概念説明があっても日本語のそれがない場合もある．その逆もある．

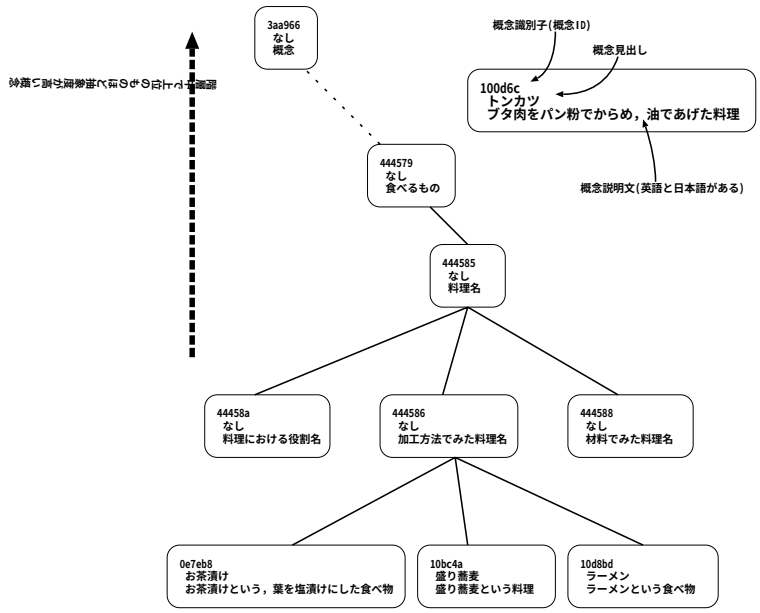

図 1.1: EDR 概念体系

1.2 概念説明の解析の必要性

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[5]．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．

[5]例えば，EDR は歴史的な文物に対する概念は豊富に収集されているが，今日的な文物に対する概念の収集には偏りがみられる．単語辞書にも同様の偏りがみられる．

EDR の概念には既に概念見出しと概念説明が付与されている．ただし，概念見出しは一意性が保証されておらず，しばしば兄弟概念間で同一の概念見出しを持つ場合があるため，この目的には十分ではない．また，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．しかし概念説明は一意であるため，もし概念説明の持つ情報が利用できるならば，概念体系から直接得られる兄弟概念の集合の中からよりその処理にとって望ましい概念を得ることが期待できる．

1.3 研究の目的

本研究は EDR 日本語概念説明を対象とし，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，について検討する．そして，それらの情報を抽出するプログラムを実装し，評価を行う．

1.4 本論文の構成

2章では，より詳細な類似概念を得るためにどのような情報が有用であるかということについて議論する．3章では，概念説明の特徴について調査結果を示す．4章では，概念説明からそれらの情報を抽出するための手法を説明する．5章では，4章で説明した手法によって目的とする情報がどれくらい得られたかを評価する．

第 2 章

概念説明の構造

本章では，EDR 概念体系を利用する際により詳細な類似概念を得るために，概念説明のどのような情報が有用であるかということを議論する．

2.1節では，従来の EDR 概念体系の利用の例を挙げ，その手法の限界を概念説明の解析によって得られる情報によってどのように補うことができるかを示す．2.2節では，2.1節で挙げた概念説明の意味的中心となる語の情報が兄弟概念間の差異を明確にする例と，そのような語の語義と，その概念説明が説明する概念の上位概念との関係に関するいくつかの例を挙げ，この情報の重要性を示す．

2.1 概念説明の情報の利用

概念説明の情報を利用した研究には，情報検索におけるクエリー拡張の研究 [5] がある．[5] では，情報検索における検索支援のためのクエリー (検索質問) の拡張をテーマとしており，クエリー拡張の手法の一つとして，概念説明の情報を利用している．しかし，EDR 概念体系から直接得られる兄弟概念を拡張クエリーとして利用するのではなく，**概念説明中にクエリーを含むような概念で拡張する**という手法で概念説明の情報を利用している．

図 2.1は，初期クエリー「航空母艦」に対して，その概念の兄弟概念で，概念説明中に初期クエリーを含むようなものを集めたものである．この場合,「カールビンソン」のような望ましい拡張 (図 2.1の○印) が得られる半面，単に概念説明文の中にクエリーを含むという条件だけでは，望ましくない拡張「制動フック」(図 2.1の×印) も得られてしまう，ということが起こりうる．

○	カールビンソン	カールビンソンという原子力 **航空母艦**
○	ニミッツ級原子力空母	ニミッツ級原子力空母という原子力 **航空母艦**
○	キエフ級空母	キエフ級**航空母艦**という **航空母艦**
○	ミンスク	ミンスクという **航空母艦**
○	ノボロシスク	ノボロシスクという **航空母艦**
×	アイランド	**航空母艦**の艦橋や砲台などを含めた 構造物
×	艦上機	**航空母艦**に積載される 飛行機
×	制動フック	**航空母艦**で, 航空機の制動に用いられる 金具

図 2.1: 初期クエリー「航空母艦」のクエリー拡張

このような問題は，**概念説明で意味的に中心的役割を担っている語の情報**がわ かっていれば，解決できると思われる．図 2.1の概念説明文の 下線部は，この概念説明において，意味的に中心的役割を担っている語を示したものであるが，この情報があれば，概念説明中に出現するクエリーが文で意味的に中心的役割の語であるのか，たまたま出現しているだけの語なのかを区別することができ，それにより,「カールビンソン」が望ましい拡張であって，且つ「制動フック」は拡張として選ぶのに不適当であるということがわかる．

更に加えて，**概念説明文の語がどのような概念で出現しているかという語義の情報**があれば，概念説明中の語でマッチングする手法に加えて，概念説明中の語の概念でもマッチングを行うことが可能となり，表層だけではマッチできなかった拡張クエリーが得られたり，表層は同じでも語義が異なる概念を拡張クエリーから除外するなど，より望ましいクエリー拡張が得られることが期待できる．

2.2 定義語

概念説明中の情報の中でも先に挙げた，**概念説明で意味的に中心的役割を担っている語の情報**は特に重要な情報であると思われる．例えば,「自動車 (0f74e9)」を親概念とする兄弟概念には

F3000(1f935e) F3000 という, レース用自動車の 分類

F3000(1f935f) F3000 という分類に属する レース用自動車

F3(1f9362) F3 という分類に属する レース用自動車

F1(1f9364) F1 という分類に属する レース用自動車

のようなものがあるが，意味的に中心的役割を担っている語 (下線部) の情報があれば，「F3000(1f935e)」が分類を表し，「F3000(1f935f)」が自動車を表す概念で，互いに異なるものを表す概念であることが機械的に区別でき，しかもこれらの概念から **F3000 に関連する概念**や**レース用自動車に関連する概念**という二通りの類似概念のグループを得ることも可能になる．

ところで，概念説明は通常の辞書における語釈文にあたるものである．辞書の語釈文を扱った研究 [4, 3, 10, 9] によると，辞書の語釈文には語釈文の意味的中心となる語が存在し，それらは定義語，中心名詞，定義動詞，などと呼ばれている．そこで本研究では，概念説明で意味的に中心的役割を担っている語のことを概念説明の**定義語**とよぶことにする．

語釈文において，定義語は通常その上位語と強い関連を持つとされている．概念説明でも通常その上位概念と強い関連を持つ．例えば「猫 (101b25)」を親概念に持つ兄弟概念では

黒猫 (0ee9f4) 黒色の毛の 猫

どら猫 (1013e8) 飼い主がなく, うろうろして食べ物などをあさる 猫

野猫 (103171) 山野に生息する 猫

下がり猫 (1ea96d) 尾が長く垂れ下がっている 猫

虎猫 (3c4dc7) 虎のような毛色の 猫

のようなものがあり，これらの定義語は全て「猫」である．例えば，これらの語義としてそれぞれ

猫 (101b25) 猫という動物

猫 (3c4fcd) 猫行火という道具

が考えられるが，上記の全ての定義語「猫」の語義は「猫 (101b25)」がふさわしい．

2.3 まとめ

概念説明には意味的に中心的役割を担っている語 (定義語と呼ぶ) が存在し，概念体系を利用する上で有用な情報であることがわかった．また，概念説明中の語の語義の情報も有用な情報である．

以上より，本研究では

- 概念説明の定義語の情報
- 概念説明中の語の語義の情報

を概念説明から抽出する．

次章ではこれらの情報を概念説明から抽出するためのヒントを得るために行った調査の結果を説明する．

第 3 章

概念説明中の手がかり語の抽出

辞書の語釈文は，あまり長くなく，文の構造が通常の文に比べてそれほど多様でないという特徴がある．そこで，辞書の語釈文を扱った研究 [10, 9, 4] では，対象の辞書の語釈文を調査し，文の構造を分類し，その構造と結びつく表層的な特徴を得ている．

本研究でも概念説明にはいくつかの特有の文の構造があると仮定した．そのような構造は特徴的な表現を伴って，概念説明中に頻繁に現れるものと思われる．したがって本研究では概念説明全体を調査対象とするのではなく，まず概念説明に頻出する表現を収集し，次に概念説明全体からそれらの表現を持つ文を選択し，それらを調査対象とする．

ある頻出表現を持つ文から特有の文の構造が見つかった場合，その頻出表現は，そのような文を解析する場合のヒント情報として用いることができる．

本研究では概念説明に頻出する表現を得るための手法として，N-gram 頻度統計を用いる．ただし，概念説明全体を対象にすると，マイナーな文にだけ通用するような強力な特徴が見逃される可能性がある．つまり，ある基準を導入して，概念説明全体をいくつかの似たような文構造を持つような概念説明の集合にグルーピングしてから，それぞれの集合の N-gram 統計をとる方が望ましいと思われる．そこで，本研究では概念説明全体の集合を**それの説明している概念を語義に持つような語の品詞**によってグルーピングする手法を提案する．

3.1節で，品詞によるグルーピングを行い，3.2節で，N-gram 頻度統計による品詞毎の特徴の調査を行う．

品詞	該当数	品詞	該当数
JAJ(形容詞)	1040	JN1;JVE(サ変動詞)	12977
JAM(形容動詞)	4614	JN2(固有名詞)	3311
JAP(連体修飾句)	18	JN3(数詞)	0
JAX(補助用言)	107	JN4(時詞)	354
JB1(接尾語)	437	JN5(前置助数詞)	12
JC1(文接続詞)	7	JN6(後置助数詞)	217
JC3(単語接続詞)	3	JN7(形式名詞)	2
JD1(普通副詞)	2468	JNM(連体詞)	150
JD2(陳述副詞)	162	JNP(体言句)	123
JEA(形容詞語尾)	0	JPR(述語句)	4092
JEM(形容動詞語尾)	0	JSE(文)	197
JEV(動詞語尾)	2	JSY(記号)	1
JIP(独立句)	190	JT1(形容詞的接頭語)	103
JIT(感動詞)	425	JT2(副詞的接頭語)	18
JJ1(助詞相当語)	0	JT3(連体詞的接頭語)	42
JJD(助動詞)	5	JT4(接頭小辞)	18
JJO(助詞)	1	JUN(単位)	241
JJP(助動詞相当語)	15	JVE(動詞)	16643
JMP(連用修飾句)	32	other(その他)	5030
JN1(普通名詞)	147435	計	200492

表 3.1: 品詞毎のグループと，該当レコード数

3.1 概念説明の品詞によるグルーピング

[3, 10, 9] では対象を特定の品詞を見出し語に持つ語釈文を対象にしている．本研究では，**その概念を語義に持つ語の品詞**によって概念説明の特徴が異なるのではないかと予想し，概念見出し辞書にある日本語概念説明を持つ概念 200492 レコードを，その概念を語義にもつ語の品詞でグルーピングした．複数の品詞をとりうるものは多数決で決定した．グルーピングの結果，名詞につく概念が圧倒的に多いことがわかった．結果を表 3.1に挙げる．

3.2 N-gram 頻度統計による品詞毎の特徴の調査

3.1で得られた，品詞でグルーピングした各々の概念説明の集合に対して N-gram 頻度統計をとり，頻出する表現 (以下，手がかり語とよぶ) を得て，それぞれが概念説明のどのような特徴を表しているのかということについて調査した．

N-gram 頻度統計をとるために，テキストファイルの N-gram を数え挙げるツール (付録 A) を実装した．

概念説明の数が 100 を割っている品詞，高頻度の N-gram が得られなかった品詞に関しては特徴的な表現はないものとして，対象としなかった．また，unigram は高頻度であっても調査の対象にはしなかった．

得られた高頻度 N-gram の中から，概念説明中の語の語義を決定するためのヒントになるもの，または，概念説明の定義語の位置を発見するためのヒントになるものがないか調査を行った．

3.2.1 文中の手がかり語

文中の手がかり語をそれぞれの品詞の高頻度 N-gram から「概念説明文の意味的構造を反映しているようなもの」を基準に選び出した．

- 主に名詞で，文中の頻出表現として「という」[1] を得た.「という」が文中に出現する概念説明は，そこで文を分割して扱うことができる．概念説明における「という」の前の部分の語は，その概念そのものと意味的に強い関連があるという特徴があり，特にこれらの語の語義を決定するに場合には，この部分を他の部分と独立に扱うことができる．

  例 「アサハン川」 アサハン川という, スマトラ島にある川

  例 「九州工業大学」 九州工業大学という国立大学

[1]形容詞 (11/1040)，形容動詞 (15/4614)，接尾語 (12/437)，普通副詞 (48/2468)，陳述副詞 (22/162)，普通名詞 (42511/147435)，サ変動詞 (30/12977)，固有名詞 (3027/3311)，時詞 (10/354)，後置助数詞 (6/217)，体言句 (13/123)，述語句 (8/4092)，文 (9/197)，単位 (177/241)，動詞 (16/16643)

- 主に名詞で，文中の頻出表現として「において」[2]「における」[3] を得た.「において」「における」が文中に出現する概念説明は，そこで文を分割して扱うことができる.概念説明における「において」「における」の前の部分は，その概念の属する分野を表しており，したがってその上位概念と意味的に強い関連があると考えられ，特にこれらの語の語義を決定する場合には，この部分を他の部分と独立に扱うことができる.

例 「相打ち」 武道などにおいて, 双方が同時に相手にダメージを与えること

例 「揚げ浜」 囲碁における上げ石

3.2.2 文末の手がかり語

概念説明において，一般には最も文末の自立語が定義語であると考えられるが，文末に頻出する表現は，そのような語ではないと考えられる．そこで，後の処理において処理対象から除く方が望ましい，文末の形式的表現 (形式 { 名詞, 動詞 }，もしくは形式 { 名詞, 動詞 } 的に用いられているような表現) を文末の高頻度語句から選び出す.

頻度の高い語句から形式名詞 (こと，さま，もの，…) や形式動詞 (する，ある) などを探し，あれば，それらを含むような頻度的にはより下位の語句も選択の対象とし，形式名詞や形式動詞的に用いている表現を手がかり語として選び出した.

3.2.3 それ以外の手がかり語

- 感動詞の「時に発する語」，後置助数詞の「を表す語」は，これらの概念説明が，概念の説明というよりは「どういう時に使う語か」を説明したものであることを表している.

例 (概念見出しが存在しない) 相手の言葉を打ち消して言い直す時に発する語

例 (概念見出しが存在しない) 心から感動した時に発する語

[2]形容動詞 (5/4614)，普通名詞 (3322/147435)，サ変動詞 (100/12977)，固有名詞 (3/3311)，後置助数詞 (4/217)，動詞 (43/16643),

[3]普通名詞 (1334/147435)，固有名詞 (6/3311)，単位 (4/241)

手がかり語	品詞と頻度
さま	形容詞 (852/1040). 形容動詞 (3484/4614), 陳述副詞 (96/162)
あるさま	形容詞 (156/1040), 形容動詞 (714/4614)
であるさま	形容詞 (131/1040), 形容動詞 (604/4614), 陳述副詞 (42/162)
様子であるさま	形容動詞 (78/4614)
気持ちであるさま	陳述副詞 (24/162)
を表す気持ちであるさま	陳述副詞 (20/162)
があるさま	形容詞 (18/1040), 形容動詞 (18/4614)
感じであるさま	形容詞 (14/1040), 形容動詞 (11/4614)
するさま	形容詞 (44/1040), 形容動詞 (152/4614)
がするさま	形容動詞 (26/4614)
感じがするさま	形容詞 (9/1040)
こと	形容動詞 (231/4614), 普通副詞 (19/2468), 独立句 (131/190) 普通名詞 (20362/147435), サ変動詞 (1956/12977), 体言句 (75/123) 述語句 (31/4092)
すること	普通副詞 (6/2466), 普通名詞 (4119/147435), サ変動詞 (631/12977) 述語句 (5/4092)
あること	形容動詞 (55/4614), 普通副詞 (7/2468), 普通名詞 (1732/147435) サ変動詞 (8/12977), 体言句 (11/123), 述語句 (4/4092)
であること	形容動詞 (45/4614), 普通副詞 (6/2468), 普通名詞 (1354/147435) サ変動詞 (5/12977), 体言句 (6/123), 述語句 (3/4092)
ということ	独立句 (52/190), 普通名詞 (92/147435), 述語句 (11/4092)
あるということ	独立句 (13/190)
であるということ	独立句 (7/190)
ものだということ	独立句 (4/190)
もの	普通名詞 (2743/147435)
あるもの	普通名詞 (81/147435)
ある	形容動詞 (175/4614), 独立句 (20/190), 普通名詞 (11/147435) サ変動詞 (14/12977), 述語句 (148/4092)
である	形容動詞 (163/4614), 独立句 (18/190), 普通名詞 (9/147435) サ変動詞 (5/12977), 述語句 (95/4092)
様子である	形容動詞 (45/4614)
ものである	独立句 (8/190)
する	サ変動詞 (4530/12977), 述語句 (717/4092)
ことをする	述語句 (11/4092)
ことができる	動詞 (2114/16463)
することができる	動詞 (327/16463)
させる	動詞 (1150/16643)
にさせる	動詞 (682/16643)
なる	サ変動詞 (265/12977)
になる	サ変動詞 (145/12977)
となる	サ変動詞 (11/12977)
している	形容動詞 (63/4614)
をしている	形容動詞 (35/4614)
様子をしている	形容動詞 (20/4614)
ようすをしている	形容動詞 (12/4614)
部分	普通名詞 (1030/147435)
の部分	普通名詞 (487/147435)
な部分	普通名詞 (28/147435)
している部分	普通名詞 (11/147435)
の一部	普通名詞 (20/147435)
状態	普通名詞 (508/147435)
の状態	普通名詞 (110/147435)
である状態	普通名詞 (15/147435)
という状態	普通名詞 (9/147435)
の一つ	普通名詞 (95/147435), 固有名詞 (28/3311)
様子	形容詞 (6/1040), 形容動詞 (33/4614), 普通副詞 (4/2468)

例 (概念見出しが存在しない) (数の後ろについて) 野球などで両チームが交互に 1 回ずつ攻撃する回数を表す語

例 (概念見出しが存在しない) (数を示す語について) 笠や笠状のものの数を表す語

3.3 まとめ

概念説明はそれが説明する概念を語義に持つ語の品詞によってグルーピングでき，各々の品詞に対する N-gram 頻度統計をとった結果，各々の品詞によって異なる特徴を持つことがわかった．

得られた手がかり語は，概念説明の定義語の位置の決定や概念説明中の語の語義を決定するのに有用であると思われる．

次章ではこれらの情報を抽出するための手法について説明する．

第 4 章

概念説明からの情報獲得

4.1 処理の流れ

本章では，概念説明の定義語の情報，概念説明中の語の語義の情報を概念説明から抽出するための手法を説明する．

まず概念説明を互いに独立に処理することのできる単位に分割する．4.2節で，概念説明を分割する手法を説明する．

分割されたそれぞれの部分文に対して，形態素解析，構文解析を行う．4.3節では形態素解析，4.4節で構文解析について説明する．

概念説明からの定義語の抽出については，4.5節で手法を説明する．

概念説明中の語の語義決定については，まず 4.6.2節で，単独で語義を決定できる語が存在することと，そのような語に対する語義決定の手法について説明する．しかし，一般には単独の語だけで語義を決定するのは困難であるので，本研究では，処理対象を個々の概念説明に限定するのではなく，意味的類似性に着目し，複数の概念説明をまたぐようなグループをつくって，グループ単位で語義を決定するという手法を用いる．4.6.1節では，そのようなグルーピングの手法を説明する．

そして最後に，デフォルトの語義決定手法について 4.6.3節で説明し，4.6.4節で，それまでに説明した語義決定の手法をまとめて，概念説明中の全ての語の語義の決定を行う手順を説明する．

4.2 概念説明の分割

3章の調査により，いくつかの手がかり語を得ているので，手がかり語の出現している概念説明については，最初に前処理を行っておく．前処理の概要は以下の通りである：

- 文末の手がかり語 (こと，さま，…) については，それを概念説明から除く[1].
- 文中の手がかり語 (という，において，における) については，そこで概念説明を分割し，個別に処理を進める．

文中の手がかり語については，それらの出現箇所で概念説明を分割して個別に扱うことにするが，どの手がかり語で分割されたかの情報が重要なので，以降「という」で分割された時の概念説明の前半の部分文字列を**という部**,「において」,「における」で分割された概念説明の前半の部分文字列を**において部**，そして残りの部分文字列 (分割されなかった場合は概念説明全体) を**メイン部**と呼ぶことにする．

例えば,「有害物質登録制度という，環境保全のための情報収集制度」の場合,「という」が含まれるので，という部は「有害物質登録制度」，メイン部は「環境保全のための情報収集制度」である．

このように可能なら分割して処理を行うことで，処理の単位を小さくし，解析の精度を高める効果が期待できる．

4.3 形態素解析

形態素解析には，JUMAN[7] を用いる．JUMAN はほとんどの場合，概念説明をうまく処理する．下記の例は，JUMAN 標準の形態素辞書にない語を JUMAN がどのように解析するかという例である (→の右側が得られた形態素区切り)：

例 「アルミノテルミー法[2]」→「アルミノテルミー法」

例 「金属酸化物[3]」→「金属 / 酸化物」

[1]文末の手がかり語について，3章では「する」という手がかり語も得ているが，これはサ変名詞に続く「する」であり，JWD はサ変動詞として見出し語の形式が「サ変名詞+する」であるために，取り除くことは行わない．

[2]「アルミノテルミー法という、金属酸化物の還元法」

[3]「アルミノテルミー法という、金属酸化物の還元法」

例 「還元法[4]」→「還元法」

例 「アルミノ珪酸塩[5]」→「アルミノ珪酸塩」

以上のように JUMAN は辞書に無い語でも内蔵のルールによって，うまく解析を行っているが，

例 「荒地野菊[6]」→「荒地 / 野菊」

例 「円錐角膜[7]」→「円錐 / 角膜」

例 「確定期限[8]」→「確定 / 期限」

例 「複式簿記[9]」→「複式 / 簿記」

のような，複合語で，語全体の語義をその語を分割した語から再構成できないような語について，その語全体が JWD には登録されているが，JUMAN 辞書には登録されていない場合 (上記例では「荒地野菊」「円錐角膜」「確定期限」「複式簿記」は分割しないでほしい) は，JWD の単語を JUMAN 辞書に追加してやらないと望ましい形態素区切りを得ることができない．

JUMAN 辞書は見出し語ベースで総数 713094 レコードある．しかし，これは見出し語の数であり，実際は「〆切」「しめ切」「しめきり」のように，一つの語に対して，そのとりうる形態を全て登録している．それらをまとめると実質はおおよそ 230000 レコードである．対して JWD は総数 395014 レコードである．

そこで，JWD の語彙と JUMAN 辞書の語彙の差分をとり，それをカスタム辞書として追加することによって，望ましい形態素区切りを得ることを試みる．その際に問題になるのは，

- JWD と JUMAN 辞書の品詞分類の違い

[4]「アルミノテルミー法という、金属酸化物の還元法」

[5]「アルミノ珪酸塩を主成分とする鉱物」

[6]「荒地野菊という植物」

[7]「円錐角膜という病気」

[8]「法律行為の履行時期を確定期限として定めた条項」

[9]「複式簿記において、帳簿上の負債や資本や収益を記す側」

- JWD と JUMAN 辞書のレコード形式の違い

である．品詞の違いについては，JWD と JUMAN の品詞で対応可能なものを調査した．結果を表 4.1に挙げる．

このような対応の下で，JWD のレコード 395014 中 384040 レコードが JUMAN 辞書と比較可能となった．

差分抽出は JWD の見出し語と JUMAN 辞書の見出し語とを比較することで行った．JUMAN 辞書に存在しない JWD の語を差分とした．得られた差分レコードの数を表 4.2 に挙げる[10]．

得られた差分レコードを JUMAN 辞書に追加した．ただし前述のように，JWD のレコード形式と JUMAN 辞書のレコード形式は異なっている．それは

- レコード内の項目の違い
- 語幹のとり方，活用型の扱い方の違い

である．レコード内項目の違いは，JWD のレコードの項目が JUMAN 辞書の要求する最低限の項目を満たしているために，項目の対応づけを行うことで容易に対応できる．しかし，JUMAN における語幹，活用型の扱いは [11] に基づいているため，EDR における語幹，活用型とは互換性がない．そのため，活用型情報を必要としない 連体詞，副詞，接続詞，感動詞，名詞，助詞，接頭辞，接尾辞については JUMAN 辞書形式に加工するのは問題なく行えた．しかし，活用型情報を必要とする形容詞，助動詞，動詞については，EDR の活用型と JUMAN の活用型に互換性がないことから，EDR の語幹や活用型の情報をそのまま使うのではなく，個別に語幹抽出と活用型推定を行わなければならない．そこで，活用型情報を必要とするこれら 3 品詞のために，[11] の規則に基づく語幹を抽出をするプログラム (付録 C) を実装し，得られた語幹の情報を用いて [11] の活用型を推定するアルゴリズム (付録 B) を開発，実装した．

以上により，差分 JWD レコードを JUMAN 辞書形式に変換し，差分をカスタム辞書として，JUMAN 標準の形態素辞書に追加し，形態素解析に用いる．

[10]JUMAN 辞書の見出し語は一部または全て漢字になっているものを含んでいるので，表の数字を引き算しただけでは差分を得られないことに注意.

JUMAN 品詞名	対応する JUMAN 辞書名	JWD 品詞
形容詞	Adj.dic	JAJ, JAM
連体詞	Adnoun.dic	JNM
副詞	Adverb.dic	JD1, JD2
判定詞	Assert.dic	
助動詞	Aux.dic	JJD, JJP
接続詞	Conjugation.dic	JC1, JC3
指示詞	Demonstrative.dic	
感動詞	Interjection.dic	JIT
普通名詞	Noun.dic	JN1
副詞的名詞	Noun.hukusi.dic	
形式名詞	Noun.keishiki.dic	JN7
固有名詞	Noun.koyuu.dic	JN2
サ変名詞	Noun.sahen.dic	JN1;JVE
数詞	Noun.suusi.dic	JN3
時相名詞	Noun.time.dic	
助詞	Postp.dic	JJO, JJ1
接頭辞	Prefix.dic	JT1, JT2, JT3, JT4, JT5
連語	Rengo.dic	
特殊	Special.dic	
接尾辞	Suffix.dic	JB1, JUN, JN6
動詞	Verb.dic	JVE

表 4.1: JWD 品詞と，JUMAN 品詞の対応表

辞書名	JUMAN レコード数	JWD レコード数	差分レコード数
Adj.dic	35881	15841	30
Adnoun.dic	1057	550	11
Adverb.dic	11077	7523	184
Aux.dic	21	333	217
Conjunction.dic	684	432	5
Interjection.dic	1498	836	69
Noun.dic	486322	271018	50683
Noun.keishiki.dic	10	104	93
Noun.koyuu.dic	33602	4667	2683
Noun.sahen.dic	60410	26504	1293
Noun.suusi.dic	37	76	20
Postp.dic	98	311	128
Prefix.dic	55	596	263
Suffix.dic	409	2457	1342
Verb.dic	78505	52792	102

表 4.2: JUMAN 辞書と JWD 辞書 (変換) と，差分

4.4 構文解析

構文解析には，KNP[6] を用いる．KNP 自体には手を加えず，そのまま用いることとする．

4.5 定義語の抽出手法

一般に，定義語は概念説明の文末の自立語である．しかし，3章の調査から，文末にはそれ自体は定義語とならないような頻出表現 (文末の手がかり語) が存在することがわかっている．そこで概念説明の定義語を，**メイン部に対する KNP の出力で，依存関係の最も底にくる文節中の自立語**[11] とする．

但しこれだけでは，一番底に並列構造があるような文では定義語を決定できない事例がある．例えば，

将校と兵士 の構文解析結果：

```
将校と<P>──┐
  兵士<P>─ PARA
```

虎と豹 の構文解析結果：

```
虎と<P>──┐
  豹<P>─ PARA
```

植物の葉と茎と根 の構文解析結果：

```
植物の────┐
 葉と<P>──┤
 茎と<P>──┤
   根<P>─ PARA
```

[11] もし一番底の文節に自立語が出現しない場合は，一つ上で定義語を探す．

のようなものである．

これらを見ると，このようなものに関しては**並列構造をなす語を全てまとめて定義語として扱う**のがふさわしいと思われる．

4.6　語義の決定手法

本節では，概念説明中の自立語の語義の決定手法について説明する．

4.6.1節では似たような語義を持つと思われる語をグルーピングする手法と，そのようなグループに属する語の語義の絞り込みと決定の手法について説明する．

但し，手がかり語や「という」部の語,「において」部の語に関しては，概念説明中において単独で語義を決定することが可能である．そこで，4.6.2節では，それらの語に対する語義決定の手法を説明する．

4.6.3節で，デフォルトの語義決定手法について説明し，4.6.4節で，これまでに説明した手法をどのようにまとめて用いるかについてまとめる．

4.6.1　グルーピング処理

概念説明は単一の文からなるため，語義を決定するための情報が一般のの文章中の語に対して不足しがちである．つまり，概念説明を独立した文として個別に扱ってしまうと，語義の決定は難しくなる．

しかし，概念説明は概念につく説明文であるから，概念体系にマッピングすることができる．そこで，概念体系の情報を個々の概念説明中の語の語義の決定に際して制約として用いることを考えた．

ある概念の兄弟概念をひとつのグループとする．そうすると，そのグループの親概念は，グループ内の概念説明の文脈の情報を表しているとみなすことができる．例えば以下は，親概念を「眼鏡 (3bdbf8) [12]」とするような概念説明のグループである：

0e276d(老眼鏡) 年をとったせいで視力が低下したときに使う眼鏡

0e69a1(遠眼鏡) 遠視を矯正するための凸レンズの眼鏡

0ed1d3(近眼鏡) 眼鏡としての近眼鏡

[12]この概念の概念説明は「眼鏡という視力調整器具」である．

0eda8e(銀縁眼鏡) 縁が銀または銀色の眼鏡

0eda90(銀縁めがね) 縁が銀色の眼鏡

0eea1a(黒縁) 黒い縁の眼鏡

3c38b9(金縁) ふちが金製または金色の眼鏡

兄弟概念は EDR 概念体系が与える意味的に最も類似した概念のグループであり，本研究では兄弟概念のなすグループをグルーピング処理における最大のグループと定義する．

概念説明中の個々の語の語義決定のためには，このグループの中に更に詳細なサブグループをつくって，グループ単位で語義の決定を行うようにする．グループには 4.6.1節で説明する **A グループ**と，4.6.1節で説明する **B グループ**がある．以降，それぞれのグループについて説明する．

A グループ

兄弟概念の作るグループ内の概念説明は同じ文脈を共有していると考えられる．したがって，それらの文に出現する同じ語は共通の語義を持つと考えられる．例えば，4.6.1節に挙げた「眼鏡 (3bdbf8)」を親概念とするグループでは,「眼鏡」による A グループを作ることができて，これらの語義は共通 (「眼鏡という視力調整器具」) である．

そこで，そのような語でグループを作る．このグループを **A グループ**と呼ぶ．A グループについては，グルーピングを行っただけではグループ内の語の語義を決定することはできないが，**A グループ内の全ての語が同じ語義を持つ**という情報を強力な意味的制約として用いることが可能となる．

B グループ

一つの文脈内で，同じ受け語に同じ格でかかっているような係り語は，似たような語義の語であると考えられる．例えば,「球技」の文脈において,「投げる」という受け語にヲ格でかかっている係り語は，どれもボールのようなものを指す語であると考えられる[13]．

そこで，**受け語と格を同じくする係り語**[14]でグループを作る．このグループを **B グループ**と呼ぶ．

[13]もちろん，中には「勝負を投げる」のような例外もあるかもしれない．

[14]構文解析 (4.4節) により，既に概念説明中の全ての係り受けの情報が得られている．

B グループにおいて係り語は，互いに似たような語義をもつと考えられる．そこで，B グループの係り文節の語ごとに語義の候補をまとめて，**複数の語の語義の候補間で親概念が共通しているものがあれば，それらの語に対しては，そのような候補をそれらの語の語義とする．**

4.6.2 単体で語義決定可能な語と，それらに対する語義決定の手法

概念説明の構造に関する情報，手がかり語の情報から，定義語,「という」部の語,「において」部の語に対しては，意味的制約を用いて語義を決定することができる．

定義語

定義語は，その上位概念と強い結び付きがあると考えられる．そこで，**語義の候補の中で，その上位概念との意味的距離の近いもの**を語義として採用するという手法が考えられる．

意味的距離はシソーラス上の概念の類似度で測ることにする．EDR 概念体系は上位下位シソーラスであるから，類似度は以下のように定義される：

$$sim(s_i, s_j) = \frac{d_c * 2}{d_i + d_j}$$

ただし，d_i, d_j は類似度を測りたい 2 つの概念 s_i, s_j のシソーラス中の深さ，d_c は 2 つの概念の共通の上位概念の深さである．

例えば，概念「相打ち」の概念説明「武道などにおいて, 双方が同時に相手にダメージを与えること」は，これまでの概念説明の解析から,「与える」が定義語であることがわかっている.「与える」のとりうる語義は以下の通りである：

支給する (0ec944)　物をあてがう
与える (1e8487)　(自分の物を) 他人に渡してその人の物とする
与える (1e8488)　(損害を) こうむらせる

これらに対して，上位概念「武道」との意味的距離を計算すると，以下のようになる：

支給する (0ec944)	物をあてがう	$\frac{1*2}{7+8} = 0.1\dot{3}$
与える (1e8487)	(自分の物を) 他人に渡してその人の物とする	$\frac{1*2}{7+8} = 0.1\dot{3}$
与える (1e8488)	(損害を) こうむらせる	$\frac{3*2}{6+8} = 0.429$

以上により「与える」の語義は,「与える (1e8488) - (損害を) こうむらせる」に決定される.

という部の自立語

という部の自立語は，その概念説明が説明する概念そのものと強い結び付きがあると考えられる．そこで，**語義の候補の中で，概念説明が説明する概念との意味的距離の近いもの**を語義として採用するという手法が考えられる．

例えば，概念 「天地人 (1f81bf)」の概念説明「天地人という，漢文の返り点」の「という」部の語「天地人」の語義の候補とその概念自身との意味的距離は以下の通りである：

天地人 (1f81bd)	宇宙の万物	$0.3\dot{6}$
天地人 (1f81be)	天地人という，三つのものの順序や順位	0.125
天地人 (1f81bf)	天地人という，漢文の返り点	1
天地人 (1f4e44)	天と地と人	0.4

以上により「天地人」の語義は,「天地人 (1f81bf) - 天地人という，漢文の返り点」に決定される.

において部の自立語

において部の自立語は，その上位概念と強い結び付きがあると考えられる．そこで，**語義の候補の中で，その上位概念との意味的距離の近いもの**を語義として採用するという手法が考えられる.

例えば，概念 「合い拳 (0e26c0)」の概念説明「拳において，両方が同じ手を出すこと」の「において」部の語「拳」の語義の候補とその親概念「引き分け (100ce0)」との意味的距離は以下の通りである：

NULL(0ef737)	握りこぶし	$0.\dot{6}$
拳 (3c3b4c)	指をつかって表す形によって勝敗を決める遊び	0.824

以上により「拳」の語義は,「拳 (3c3b4c) - 指をつかって表す形によって勝敗を決める遊び」に決定される.

4.6.3 デフォルトの語義決定手法

定義語でない語，特別のヒントの情報のない文節の語に対するデフォルトの語義決定の手法について説明する．

語義の決定は，単一の語に対して試みるのではなく，係り受け関係にある語の組に対して試みる (構文解析 (4.4節) により，全ての語の間の係り受け関係が得られている)．

そこで，**係り受けの関係にある 2 語に対するスコアを計算し，スコアの最も高かった語義の候補の組合せを 2 語の語義とする**．

スコア

係り受けの関係にある 2 語 (w_g, w_d) に対し，とりうる概念の組 (s_g, s_d) が考えられるとき，スコア $Score(w_g, w_d, s_g, s_d)$ を以下のように定義する：

$$Score(s_g, s_d) = \frac{f(s_g, s_d)}{f(s_g) * f(s_d)}$$

$Score(s_g, s_d)$ は Mutual Information に基づくスコアで，2 つの概念の共起しやすさを評価している．

ある語の組 w_g, w_d に対して複数の語義が考えられる場合，その全ての組合せに関して図 4.6.3のスコアを計算し，比較を行い，最も高いスコアを出した語義の組合せを正解として採用する．語や概念の頻度情報は，EDR 共起辞書 (以下，JCC と呼ぶ) から抽出することができる．

sparseness の解消

このスコアは JCC から得られる共起頻度に基づいているため，例えば正解の語義のの組に対して，たまたま JCC から共起頻度が得られなかった場合，その組に対するスコアの値は 0 になってしまう．このような JCC の sparseness の解消のために，概念クラス [8] を導入する．概念体系中のある概念に対する概念クラスとは，**その概念の下位概念全て**である．

本手法では，Mutual Information に基づく本スコア中の，かかり語に対して概念クラスによる拡張を適用する．概念クラス C を適用したスコアの式は，以下のようになる：

$$Score(s_g, C_{s_d}) = \frac{f(s_g, C_{s_d})}{f(s_g) * f(C_{s_d})}$$

ただし，

$$f(s_g, C_{s_d}) = \sum_{s \in C_{s_d}} f(s_g, s), f(C_{s_d}) = \sum_{s \in C_{s_d}} f(s)$$

共起頻度を得る対象を概念クラス全体に広げることで，ある程度 sparseness は解消できると考えられる．しかし，対象となる概念が概念体系における葉ノードだった場合などは，概念クラスは自身の概念しか含まないために，これでは十分ではない．そこで，概念クラス中の全ての概念について共起頻度が得られなかった場合は，**共起頻度が得られるまでその上位概念クラスを辿ってスコアを計算する**という方法で sparseness の解消をはかる．

事前の語義の候補の絞り込み

このスコアに基づく語義決定の手法は，構文解析の結果，全ての文節間の係り受け関係を得ることができているので[15]，デフォルトの手法として利用できる．但しこのままでは，

- 全ての語義候補の組み合わせに対してスコアを計算するため，計算量が莫大となる
- 概念体系の情報が使われていない

という問題がある．

そこで，語義の候補 (の組み合わせ) に対してスコア計算を行う前に，**それらの概念説明が属するグループの親概念との意味的距離を評価し，他の候補のそれに比べて明らかに距離的に離れている候補を取り除く**という処理を行う．これは概念体系の情報を使った絞り込みの処理である．

具体的にはそれぞれの語義の候補に対して親概念との意味的距離 d を求め，これら候補間での平均 $\overline{d}$ に対して

$$\overline{d} > d$$

となる d を値としてとるような候補を除くという処理を行う．

この処理によって，計算量の爆発をある程度抑える効果が期待できる．またこの処理は，候補の中から現在の文脈に対して意味的に遠い候補を除く処理であるから，不正解の候補であるにも関わらず，たまたま JCC で高い頻度を持っているために，スコアリングで正解として選ばれてしまうことを防ぐ効果が期待される．

[15]それだけでなく，受け文節を共有するような複数の係り受けがあるような場合では，スコアを組み合わせて制約を加えることもできる．

4.6.4 概念説明中の語の語義決定の手法 (まとめ)

これまで説明した語義決定の手法を組み合わせて，概念説明中の全ての語の語義をどのように行うかについて以下にまとめる：

1. まず，親概念を同じくするグループを作り (4.6.1節)，これを処理の最大の単位とする．
2. 次に兄弟概念のグループ内で，可能な A グループ (4.6.1 節)，B グループ (4.6.1節) を全てまとめあげる．
3. 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，4.6.2節の手法に従い語義を決定する．
4. 3で語義の決定したものが A グループにあれば，グループ内の全ての語の語義を fix する．
5. B グループの受け語で，これまででまだ語義の決まっていないものに対して，4.6.1節で説明した手法に基づき，語義の決定を行う．
6. この段階で語義の決まっていない語を含んでいる全ての係り受けの組に対してデフォルトの語義決定手法 (4.6.3節) を適用し，語義の決定を行う．係り受けの組内で既に語義の fix された語に関しては，全ての組合せを計算するのではなく，その語義を含む組合せのみでスコアを計算する．

第 5 章

評価

4章で述べた手法の有効性を評価するために実験を行った.

5.2節では，カスタム辞書追加による概念説明に対する形態素解析の精度の変化について評価する.

5.3節では，定義語抽出の手法の精度を評価する.

5.4.1節では，定義語の語義の決定手法の精度を評価する.

5.4.2節では,「という部」の語の語義の決定手法の精度を評価する.

5.4.3節では,「において部」の語の語義の決定手法の精度を評価する.

5.5節では，A グループの意味的制約の強さを評価する.

5.6節では，A グループの情報を定義語,「という」部の語,「において」部の語の語義決定に利用し，精度が向上したかどうかを評価する.

5.7節では，B グループの語の語義決定手法の精度を評価する.

5.8節では，デフォルトの語義決定手法の精度を評価する.

5.9節では，いくつかの兄弟概念グループに対して，全ての語義決定処理を行い，全体的な語義決定の評価を行う.

5.1 評価セット

本研究で提案している語義決定手法を評価するために，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.

5.2 形態素解析

概念説明からランダムに選択した 100 文で解析精度を調べた．文中の全ての形態素区切りを正しく得た時に限り正解とすると，juman 標準辞書のみでも 85 文正解したが，さらに形態素辞書を追加すると，97 文正解するようになった.

標準辞書で失敗して，追加辞書でうまくいった場合 (11 文) の全てが，上記例のように，辞書の追加によって切り過ぎを押えることによって望ましい形態素区切りを獲得したものだった.

追加辞書で失敗した 3 文は以下のようなものであった：

- **いちかばちかで目的を果たそうとする**
  - 「いちかばちか」を「い / ちかば / ちか」と区切ってしまう (両方とも同じ結果).「いちかばちか」は juman 標準辞書にも JWD にも登録されていない.
- **船舶やホテルなどの休憩室**
  - 「休憩室」を「休 / 憩室[1]」と区切ってしまう (両方とも同じ結果).「休憩」というサ変名詞は juman 標準辞書，JWD 両方に登録されている.
- **同じことをくどくど言い続ける**
  - 「くどくど」を「くど / くど[2]」に区切ってしまう (両方とも同じ結果).「くどくど (副詞)」は juman 標準辞書にも JWD にも登録されていない.

このうち「休憩室」は JUMAN の形態素重みのパラメータを操作することで正しく解析できるようになるが，パラメータの変更は元々正しく解析されていた文に影響を与える可能性があるので，本研究ではパラメータを操作しない.

[1]管状の臓器に見られる, 憩室という異常な突起

[2]「火をたいて食物を煮たきするように造られた, 土やれんが製などの設備」もしくは「かまどの後ろにつけた煙を出す穴」

5.3 定義語の抽出手法の評価

4.5節の定義語の定義で取り扱えないような概念説明がないかどうか，概念説明 535 文に対して確認を行った．その結果，定義語の抽出に失敗した概念説明は 13 文あった．

うち 12 例は，形態素解析の段階での失敗であった：

- 形態素解析に失敗．JWD から新たに追加した語彙のせいで，正しくない形態素区切りを得てしまった事例 (2 例)

  北方 (0ec107) 「北の**方角**」を「北の方[3]/角」

- 形態素解析の失敗．語彙の不足 (JUMAN 辞書にも JWD にもない語彙) の事例 (1 例)

  特別区 (1f39a7) 「東京都の **23 区**」を「23/区」

- 形態素解析の失敗．形態素区切り自体は正しく得られているが，それらの形態素が表層は同じだが異なる語として認識されている事例 (9 例)

  北部 (108a77) 「北の**部分**」で,「部分」を「ぶわけ」という語として認識.

しかし，1 例においては本研究で提案した定義語抽出の手法の枠組では扱えないものであった：

- 「という」が 2 回現れ，且つ 2 つの「という」が同格であるような事例.

  戸障子 (100643) 「雨戸**という**建具と障子**という**建具」

このようなものが他にもあるのかどうかについて，全概念説明を調査したところ，全部で 119 文存在することがわかった．但しこれらには 2 つのパターンがある．一つは，上記の例のように 2 つの「という」が同格であるようなものだが，もう一つは

金太郎 (0ed3c1) 「金太郎**という**，足柄山で山姥に育てられた**という**よく太って体の赤い怪童」

桂 (0ea091) 「月に生えている**という**，桂という想像上の植物」

[3]「北の方 (きたのかた)」という名詞

のように，どちらかの「という」にだけ着目すればいいようなものである．

「戸障子」のような例は，先に構文の情報を持っていないとわからないもので，本研究で提案した手法の枠組を越えたものである．一方,「金太郎」や「桂」のような例は，前方か後方どちらかの「という」にのみ着目すれば対応できる．119 文を観察したところ，両方の「という」が同格であるものが 42 文，前方の「という」にだけ着目すればよいものが 62 文，後方の「という」にだけ着目すればよいものが 12 文，概念説明の記載ミスと思われるものが 1 文[4]だった．

以下の評価では，ad hoc な対応として，前方の手がかり語のみに着目して,「という」部,「において」部を抽出した．

5.4 単体で語義決定可能な語に対する語義決定の手法の評価

5.4.1 定義語の語義決定実験結果

4.5節の手法により，全 535 文から 547 の定義語が得られた．これらに対し，4.6.2節の語義決定手法を適用し，精度を調査した．

この手法はシソーラス上の意味的距離をスコアにしているので同じ値が出やすく，一位のスコアで正解を決定しても，複数の正解をとりうる．そして，正解もまた一般には一意でない．そこで，一語毎に真の正解語義集合とシステムの出した正解語義集合との比較を行い recall と precision を算出し，その平均で評価を行った．結果を表 5.1に挙げる．

$$Recall = \frac{\text{システムが出した真の正解}}{\text{真の正解}}$$

$$Precision = \frac{\text{システムが出した真の正解}}{\text{システムが正解としたもの}}$$

5.4.2 「という部」の語の語義決定実験結果

3.2節で得た文中の手がかり語の情報を用いて，全 535 文から 118 の「という」部が得られ，その中から 143 の自立語が得られた．これらに対し，4.6.2節の語義決定手法を適用し，精度を調査した．

[4]**中世 (3c0837)** の,「古代と近世の間の中世**というという**時代区分」

評価対象総数	547
対象の語が JWD に存在しなかったために語義の候補が得られなかった数	6
語義の候補が一つしかなかった数	237
語義の候補が全て概念体系上に存在しなかったために語義が決定できなかった数	0
語義の候補の中に真の正解が存在しなかった数	13
評価対象実数	291
語義候補数の平均	4.237
システムの出した正解語義数の平均	0.268
真の正解語義数の平均	1.612
recall の平均	0.540
precision の平均	0.636

表 5.1: 定義語の語義決定精度

この手法も定義語と同じく，それぞれの語毎に真の正解語義集合とシステムの出した正解語義集合との比較を行い recall と precision を算出し，その平均で評価を行った．結果を表 5.2に挙げる．

5.4.3 「において部」の語の語義決定実験結果

3.2節で得た文中の手がかり語の情報を用いて，全 535 文から 18 の「において」部が得られ，その中から 21 の自立語が得られた．これらに対し，4.6.2節の語義決定手法を適用し，精度を調査した．

この手法も定義語と同じく，それぞれの語毎に真の正解語義集合とシステムの出した正解語義集合との比較を行い recall と precision を算出し，その平均で評価を行った．結果を表 5.3に挙げる．

5.4.4 意味的距離に基づく語義決定の問題点

定義語や,「という」部の語,「において」の部の語のほとんどに対して，意味的距離によるスコアリングは高い精度で語義を決定できることがわかった．

しかしこのスコアでは，対象となる語が非常に抽象的な意味だった場合に正しく語義を決定することができない．なぜならこのような時，正解の語義候補は抽象的な概念なの

評価対象総数	143
対象の語が JWD に存在しなかったために語義の候補が得られなかった数	10
語義の候補が一つしかなかった数	81
語義の候補が全て概念体系上に存在しなかったために語義が決定できなかった数	0
語義の候補の中に真の正解が存在しなかった数	2
評価対象実数	50
語義候補数の平均	3.36
システムの出した正解語義数の平均	0.62
真の正解語義数の平均	1.4
recall の平均	0.834
precision の平均	0.808

表 5.2: 「という」部の自立語の語義決定精度

評価対象総数	21
対象の語が JWD に存在しなかったために語義の候補が得られなかった数	0
語義の候補が一つしかなかった数	15
語義の候補が全て概念体系上に存在しなかったために語義が決定できなかった数	0
語義の候補の中に真の正解が存在しなかった数	0
評価対象実数	6
語義候補数の平均	2.5
システムの出した正解語義数の平均	0.5
真の正解語義数の平均	1.833
recall の平均	0.667
precision の平均	0.75

表 5.3: 「において」部の自立語の語義決定精度

で，概念体系中のかなり上の階層に位置していて，より葉ノードに近いその他の候補に比べて不利なスコアになってしまうからである．

例えば「腹合わせ (1043d2)」の概念説明「共同して事をすること」中の語「事」に対しては，以下のような語義が考えられる：

概念識別子	概念見出し，概念説明	概念体系中の階層
3d1815	事 ある事柄に関して言えば	11 段目
3ce7f2	任務 自分の責任において遂行する事柄	8 段目
3d017c	物事 ものごと	2 段目
3d017d	事象 事象	4 段目
0ed533	事 ある事物に関して	9 段目
3cf180	ありさま 物事の状態	7 段目

この場合「事」は非常に抽象的な意味で用いられていて，3d017c がふさわしいと考えられる．しかし，ここで各々の概念体系中におけるトップノードからの深さを見てみると，3d017c は 2 段目であり[5]，他に比べて極端に不利であることがわかる．

しかし，抽象的な意味をもって出現している語は概念説明でそれほど重要な役割を担っているとは考えにくく，このような語の語義の決定は優先して行う必要はないと考えられる．

5.5 A グループの意味的制約の評価

評価セット全 535 文から 222 の A グループが得られた．これらの A グループには 812 語の語が含まれていた．これらの語に対し，実際に語義を人手で決定し，A グループ内でそれらの語義が一致するかを調査した．

その結果，56 グループ 112 語については語義が一致しなかった．このような事例は「という」部の語で，且つ語義決定の際にスコアが最大値 (1) で決定されたものに対して起こっていた．

例 親概念「門戸 (10bcfa)」中の「貴人口」という A グループ：

[5]親概念「連袂する (10e926)」の深さは 7 段目

貴人口 (0ec2dd) 能舞台の貴人口という戸口

貴人口 (0ec2de) 茶室において, 貴人口という出入口

この例のように,「という」部の語はその概念説明の説明する概念そのものと意味的に最も近く，実際このような場合は，両者の語義は等しく扱うべきではないと考えられる．

そこで，このようにシソーラス内の意味的距離が最大値 (1) で決定された語義に関しては，A グループよりもさらに強い意味的制約が加わっているものとして，強制的に語義を一致させるような処理は行うべきではないと思われる．

5.6 定義語らに対する A グループの情報の利用についての評価

5.5節の調査により，A グループの情報は語義を決定する際に有効に利用できることがわかった．そこで，定義語,「という」部の自立語,「において」部の自立語の語義を決定する際に，これら A グループの情報を利用した．具体的には，前節の実験でシステムが出力した正解語義集合の中で，おなじ A グループに属している語があったら，それらの正解語義集合を比較し，一致する語義だけを残すという絞り込み処理を行った[6]．

しかし，定義語,「という」部の自立語,「において」部の自立語のどれも，A グループ中の語同士で正解語義集合が完全に一致したため絞り込みは行うことができなかった．これは，定義語や「という」部の自立語,「において」部の自立語の中の A グループの語で，複数箇所 (メイン部,「という」部,「において」部) をまたがって出現したものが一つも存在しなかったことを意味[7]し,「同じ ような役割を持った語が概念説明中で同じ ような箇所に出現する」という本来の仮定を支持する結果である．

A グループの情報は，以降の語義決定の処理でも利用する．

[6] 5.5節で挙げたように，スコアが最大値で決定されたものに対してはこの処理は行わない．

[7] 例えば，もし全て定義語で出現していれば，語義の決定の手法は全て一致するので，全く同一の正解語義集合しか得られない．

評価対象総数	368
多義でなかった数	96
語義の候補の中に真の正解が存在しなかった数	8
評価対象実数	264
語義候補数の平均	4.17
システムの出した正解語義数の平均	3.31
システムの出した正解語義数の平均 (A グループ情報利用)	3.12
真の正解語義数の平均	1.40
全ての候補を落してしまった数	35
全ての候補を残してしまった数	211
recall の平均	0.81
precision の平均	0.58
全ての候補を落してしまった数 (A グループ情報利用)	39
全ての候補を残してしまった数 (A グループ情報利用)	225
recall の平均 (A グループ情報利用)	0.85
precision の平均 (A グループ情報利用)	0.62

表 5.4: B グループの自立語の語義決定精度

5.7 B グループの語の語義決定手法の評価

4.6.1節で述べた手法により，全 535 文から 125 の B グループが得られ，その中から 372 語の自立語が得られた．これら B グループの語に対し，4.6.1節で述べた手法を適用し，語義を決定した．ただし，これらの語の中に定義語,「という」部の語,「において」部の語, A グループの語が含まれていて，既に語義が決定していた場合は，それらの語の語義の情報は利用した．既に語義の決定していた語を除くと，評価対象の総数は 368 語となった．

この手法も定義語と同じく，それぞれの語毎に真の正解語義集合とシステムの出した正解語義集合との比較を行い recall と precision を算出し，その平均で評価を行った．結果を表 5.4に挙げる．

高い recall 値の原因は評価対象で，語義候補全てが残ってしまう事例が多かったことによる．また，A グループの情報により，recall，precision 値はそれぞれわずかに高くなった．

評価対象総数	605
対象の語が JWD に存在しなかった数	176
語義の候補が一つしかなかった数	188
語義の候補の中に真の正解が存在しなかった数	10
評価対象実数	231
語義候補数の平均	5.12
システムの出した正解語義数の平均	1.35
システムの出した正解語義数の平均 (絞り込みを行った)	1.21
真の正解語義数の平均	1.70
recall の平均	0.33
precision の平均	0.37
recall の平均 (絞り込みを行った)	0.27
precision の平均 (絞り込みを行った)	0.29

表 5.5: デフォルトの語義決定精度

5.8 デフォルトの語義決定手法の評価

これまでの語義決定の処理の対象にならなかった概念説明中の 605 の自立語に対し，Mutual Information に基づくデフォルトの語義決定手法を適用した．

対象となる語を含む係り受けの組のうちどちらかが既に語義が決定していた場合は，その語義との組合せでだけスコアリングを行った．また，対象となる語が A グループに含まれていて語義が決定できる場合は，その語義を正解とした．

語義決定実験は，4.6.3節で説明した事前の候補の絞り込み処理を行わなかった場合と行った場合の二通りで行った．絞り込み処理を行った実験は絞り込み処理を行わない実験に比べ，約 3 分の 1 の時間で処理を終えた[8]．デフォルトの語義決定実験の結果を表 5.5に挙げる．

実験の結果，このような絞り込みを行うと，かえって語義決定の精度が低くなってしまうことがわかった．これは，このような概念体系中の意味的距離に基づく絞り込みが，これらの語の正解語義を落してしまっているということであり．定義語,「という」部の語,

[8]OS: Solaris 2.6, CPU: Pentium II * 2, Memory: 512MByte のマシンで，絞り込み処理を行った実験は約 2 時間，絞り込み処理を行わない実験は約 6 時間かかった．

「において」部の語のような,『その概念説明が説明する概念やその概念の親概念と強い関連を持つような語』以外の語では，概念体系の情報を用いた意味的制約は有効に働かないということを示している.

また，全ての語義の候補について共起辞書から頻度が得られなかった語が 104 語あった.

5.9 語義決定の全体的評価

評価セット中の全評価対象に対する語義決定精度をまとめたものを表 5.6に挙げる．ただし，B グループの語義決定では A グループの情報を利用したもの，デフォルトの語義決定手法が適用された分については，事前の語義候補の絞り込み処理を行わなかったものの結果でまとめた.

評価対象総数	842
recall の平均	0.60
precision の平均	0.55

表 5.6: 全体の語義決定精度

第 6 章

結論

一般に，辞書の語釈文はあまり長くなく，文の構造が通常の文に比べてそれほど多様ではないという特徴がある．本研究では，EDR の概念説明にもいくつかの特有の文の構造があると仮定し，そのような構造を反映した表現を得るために，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．

そのような頻出表現を手がかりに，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」部と「において」部を定義した．

形態素解析では，EDR 日本語単語辞書の語彙を形態素解析器の辞書に変換し追加することで概念説明に対する形態素解析精度を向上することができた．

意味解析では，概念説明中の自立語に対して語義の決定を行った．定義語,「という」部の語,「において」部の語に対しては

- 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，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の決定を行うことができた．

第 7 章

今後の課題

- 本研究では，形態素解析や構文解析の前に文末手がかり語の除去や文中手がかり語による文の分割などの前処理を行い，一定の効果を得た．

  しかし，概念説明全体からみて量は少ないものの， 5.3節にあるように本研究で仮定した文の構造では対応できない概念説明が存在した．このような概念説明では，手がかり語に関する処理を行う前に構文の情報が得られている必要がある．

  各解析の適用順序を見直すことで，このような事例を正しく取り扱うことが可能になると思われる．

- 本研究では意味解析において各種のスコアリングによって語義決定を行ったが，全てスコアが 1 位のものだけを採用し，語義としていた．

  しかし，語によっては人間による判断においても，単一の概念を語義として決めるのは難しく，多数の概念を語義として選ばざるをえなくなることことも珍しくなかった．そのような語に対してスコアが 1 位のものだけを語義として選んでいると，recall が低くなってしまうということが起こっていた．

  そこで，この採用の順位を 2 位以下まで広げることで，比較的低い recall 値を改善できる可能性がある．

- 定義語,「という」部の語,「において」部の語の語義決定に用いたスコアは，高度に抽象的な語義を持つような語 (「何か」,「事」など) に対して不利なスコアであった．

  しかし実際の概念説明を見ると，そのように抽象的に用いられている語は，本来概

念説明全体からみてそれほど重要な役割を担っているとは考えにくく，このような語の語義決定を行うことの重要性は，それ以外の語のそれに比べて低いのではないかと考えられる．

そこで，そのような語をあらかじめフィルタリングすることが考えられる．そのような語は形式 { 名詞，動詞 } 的に用いられていて，他の具体的に使われている語に比べて比較的高頻度に出現していると考えられるので，このような語のリストアップにも N-gram 頻度統計が利用できるかもしれない．

- B グループの語の語義決定について，意味的に近い概念をまとめるために兄弟概念でまとめるという手法を用いたが，他にも，例えば意味的距離によるクラスタリングなどのような手法なども考えられる．

- 手法全体のカバレジは約 32%程度だが，これは主に EDR 辞書内における consistency の問題 (EDR 単語辞書に語が存在しないこと，EDR 概念体系に語義の候補の概念が存在しないこと) と，EDR 共起辞書の sparseness の問題 (共起辞書に共起頻度が存在しないこと) によるものであり，EDR 辞書の拡充により，解決される問題である．

謝辞

本研究を進めるにあたり，終始御指導くださいました奥村学助教授に心から感謝致します.

常日頃より研究に関して数多くのアドバイスをくださいました島津明教授，大石亨助手，そして自然言語処理学講座の皆様に感謝致します.

EDR parsing library は本研究の実装を行うにあたって，欠くことのできないライブラリでした．作者の本田岳夫氏と徳田昌晃氏に感謝致します.

最後に，JAIST での 3 年間を物心両面にわたって支援してくださった家族と友人に感謝致します.

参考文献

[1] Makoto NAGAO and Shinsuke MORI. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In *COLING 94*, pp. 611–615, 1994.

[2] (株) 日本電子化辞書研究所. EDR 電子化辞書使用説明書, 第 2 版, March 1995.

[3] 酒井桂一, 中村順一, 長尾真. オンライン辞書定義文の解析と知識ベース化. 情報処理学会研究報告 89-NL-71, pp. 1–8, 1989.

[4] 鶴丸弘昭, 竹下克典, 伊丹克企, 柳川俊英, 吉田将. 国語辞典情報を用いたシソーラスの作成について. 情報処理学会研究報告 91-NL-83, pp. 121–128, 1991.

[5] 太田千晶. 電子化辞書を利用した, 概念に基づくクエリーの拡張に関する研究. Master's thesis, 北陸先端科学技術大学院大学, 1997.

[6] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6, June 1998.

[7] 黒橋禎夫, 長尾真. 日本語形態素解析システム JUMAN version 3.5, March 1998.

[8] 相場徹. 構文・意味解析と統合した形態素解析に関する研究. Master's thesis, 北陸先端科学技術大学院大学, 1995.

[9] 冨浦洋一, 山尾昭博, 日高達, 吉田将. 語義文に表現されている動詞間の上位－下位関係－機能語列の取り扱いについて－. 情報処理学会研究報告 91-NL-81, pp. 33–40, 1991.

[10] 冨浦洋一, 日高達, 吉田将. 国語辞典の語義文からの動詞の上位－下位関係の抽出. 情報処理学会研究報告 89-NL-73, pp. 17–24, 1989.

[11] 益岡隆志, 田窪行則. 基礎日本語文法. くろしお出版, 1989.

第 A 章

N-gram 頻度統計ツール ngram

ngram はテキストファイルの N-gram 頻度統計をとるツールである．[1] のアルゴリズムに基づく実装である．C 言語で実装されているため，ほとんどのプラットフォームで動作する．以下のような特徴を持っている：

- 単語単位でも文字単位でも N-gram 頻度統計をとることができる．
- 任意の N に対して N-gram を得ることができる．
- 対象となるテキストファイルの大きさに制限がない．
- テンポラリファイルをつくって計算するので，非常に少ないメモリで動作する．

また，ngram は日本語テキストファイルの N-gram 頻度統計をとることができる．日本語の処理に関して，以下の特徴を持っている：

- 入力テキストファイルの文字エンコーディングは，日本語 EUC と，ISO-2022-JP を利用可能．
- 入力テキストファイルの文字エンコーディングを自動で判定する．
- 入力テキストファイル中の，(いわゆる) 半角英数字と，(いわゆる) 全角英数字を同一視して数えることが可能．

第 B 章

活用型情報の必要な品詞に対する活用型推定

各品詞毎に以下のアルゴリズムで活用型を推定した．但し，EDR と JUMAN では語幹の扱いが異なる (付録 C) ことも考慮しなければならない．すなわち，JWD の語幹の情報は利用できない．

B.1 動詞

if (EDR の活用型が JRV4)
  カ変動詞
else
  if (語幹が母音で終わる)
    母音動詞
  else
    if (語幹が 'k' で終わる)
      子音動詞カ行
    else if (語幹が 'g' で終わる)
      子音動詞ガ行
    else if (語幹が 's' で終わる)
      子音動詞サ行

else if (語幹が 't' で終わる)
子音動詞タ行
else if (語幹が 'n' で終わる)
子音動詞ナ行
else if (語幹が 'b' で終わる)
子音動詞バ行
else if (語幹が 'm' で終わる)
子音動詞マ行
else if (語幹が 'r' で終わる)
if (EDR の活用型が JRV7)
子音動詞ラ行イ形
else
子音動詞ラ行
else if (語幹が 'w' で終わる)
子音動詞ワ行

- サ変動詞はサ変名詞にしているので，対象にしなくてよい.
- JRV4 はカ行変格活用，JRV7 はオッシャル活用 (連用形に「い」をとりうる) を表す.

B.2 形容詞

if (EDR の品詞が JAJ)
if (語幹が 'a' で終わる)||(語幹が 'u' で終わる)||(語幹が 'o' で終わる)
イ形容詞アウオ段
else if (語幹が 'i' で終わる)
イ形容詞イ段
else if (EDR の品詞が JAM)
if (EDR の活用型が JRM7)||(EDR の活用型が JRM8)
タル活用

else

if (EDR の活用型が JRM1)||(EDR の活用型が JRM2)
||(EDR の活用型が JRM5)||(EDR の活用型が JRM9)
ナ形容詞

else if (EDR の活用型が JRM3)||(EDR の活用型が JRM4)||(EDR の活用型が JRMA)
ナノ形容詞

- JAJ は形容詞，JAM は形容動詞を表す.
- JUMAN 辞書で実際にイ形容詞イ段特殊をとるのは「おっきい」「おおきい」のみ.
- 連体形に「な」をとるものがナ形容詞で，さらに「の」もとれるものがナノ形容詞.

B.3 助動詞

if (EDR の活用型が JRA1)||(EDR の活用型が JRA2)||(EDR の活用型が JRA3)
if (語幹が 'i' で終わる)
イ形容詞イ段
else if(EDR の活用型が JRM1)||(EDR の活用型が JRM2)
||(EDR の活用型が JRM5)||(EDR の活用型が JRM9)
ナ形容詞
else
無活用型

第 C 章

EDR と JUMAN の語幹の違いについて

例えば「貸す」を例にとると，EDR では「貸」を語幹 (不変化部) とするが，JUMAN で採用している文法 [11] では語幹は「kas」(基本系語幹[1]) である．

このため，EDR の語幹情報は JUMAN 辞書トランスレートには利用できない．そこで，

1. JWD 単語の終止形をローマ字表記にする．

2. 末尾の「u，ru」を取り除く ([11]p.16 参照)

の処理を行って語幹を新たに抽出しなければならない．今回ローマ字表記を得るために，フリーソフトウェアの kakasi を利用した．

[1][11] では，結合する活用語尾の違いに応じて「基本系語幹」と「タ系語幹」の 2 つがある．この場合のタ系語幹は「kasi」