JAIST Repository
https://dspace.jaist.ac.jp/

Title	EDR概念辞書とコーパスを用いた語義曖昧性解消
Author(s)	八木，恒和
Citation	
Issue Date	2004-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1888
Rights	
Description	Supervisor:白井 清昭，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

EDR 概念辞書とコーパスを用いた語義曖昧性解消に関する研究

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

八木 恒和

2004 年 9 月

修 士 論 文

EDR 概念辞書とコーパスを用いた語義曖昧性解消に関する研究

指導教官　白井 清昭

審査委員主査　白井 清昭 助教授
審査委員　島津 明 教授
審査委員　鳥澤 健太郎助教授

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

210096 八木 恒和

提出年月: 2004 年 8 月

Copyright © 2004 by Tsunekazu Yagi

概 要

自然言語処理を行う際の重要な問題の一つに語義曖昧性解消 (WSD) がある。語義曖昧性解消の過去の研究として教師あり学習がよく行われている。しかし教師あり学習は正解付きデータを必要とし、正解付きデータの中で良く出現する単語に関しては良い結果が得られるが、低頻度語については学習が困難である。また本研究は、人間の文章理解を支援する読解支援システムでの使用を前提としているが、このような場合は低頻度語を含めた多くの単語に対して語義曖昧性解消を行う手法が求められる。この問題を解決するために、語義タグ付きコーパスと辞書定義文を利用した分類器を作成する手法を提案する。そして、高頻度語には教師あり機械学習の手法を使用し、低頻度語には辞書の定義文を用いた分類器を使用する。この 2 つを組み合わせることにより頑健な WSD システムの構築を行う。

まず、辞書定義文を用いた分類器について説明する。例えば、「犬」という単語があり、その語義として「犬という動物」と「スパイという役割の人」の 2 つがあるとする。そして、「犬にえさをあげる」という文の「犬」の語義を判定する場合を考える。「犬」が語義タグ付きコーパスに一度も出現しない場合は、「犬」の語義を判定するモデルを学習することはできない。ここで辞書定義文に含まれる語義の上位概念に着目する。語義の上位概念は辞書定義文の末尾から取出せる。例えば、「犬という動物」から「動物」という上位概念が取出すことできる。ここで、コーパスに「象にえさをあげる」「猫にえさをあげる」などの文があり、象、猫の上位概念は「動物」であるとする。このとき上位概念「動物」と「えさ」が共起することを学習できるので、「犬にえさをあげる」という文では「犬」の語義が「犬という動物」であることがわかる。このように、辞書定義文から語義の上位概念を抽出し、上位概念と周辺語との共起性を学習することにより、低頻度語でも単語の語義を正確に判定できる。

次に、上位概念を用いて語義曖昧性解消を行う手法について説明する。ここでは上位概念を用いた Naive Bayes モデル $P(s)\prod_i P(f_i|c)$ を学習し、その確率が最大となる語義を選択する。この式において、f_i は素性、c は上位概念、s は語義を表わしている。また、語義を決めたい単語 (w) の直前・直後に存在する単語の表記や品詞、w の前後 20 単語以内に現れる自立語の基本形、w と係り受け関係にある単語の基本形などを素性 f_i として使用する。

次に、辞書定義文から上位概念を抽出する手法について述べる。基本的には、辞書定義文の末尾にある単語を上位概念として取り出す。また、上位概念が辞書定義文の末尾以外になるときもある。そのような場合を考慮して抽出パターンを作成した。例えば、「欠点をさがしだして言う悪口のこと」というような「N のこと」で終わる辞書定義文から N(この場合は「悪口」) を取出すパターンを作成した。64 個の抽出パターンを作成し、EDR 概念辞書の辞書定義文から上位概念を抽出した。上位概念を抽出できた語義数を EDR 概念

辞書中の全ての語義数で割った値は98%であった。また、抽出した上位概念をランダムで200個選択し、人手で判定したところ、185個が適切な上位概念であった。

次に、高頻度語のWSDを行う教師あり機械学習の手法について説明する。ここでは、教師あり機械学習アルゴリズムとして、Support Vector Machine(SVM)を使用する。SVMで使用した素性として、Naive Bayesモデルで使用した素性以外に、wの二つ前または二つ後の単語または品詞、wの直前または直後に現れる2つの単語の表記または品詞の組を追加した。

本研究では、最終的に高頻度語のためのSVMによる分類器と低頻度語のための上位概念を用いた分類器の2つを組み合わせた。組みあわせる手法は以下の通りである。語義を決めたい単語の訓練データにおける出現頻度がある閾値以上ならSVMによる分類器を、それ以外はNaive Bayesモデルを用いて判別を行う。本研究ではこの閾値を5とした。

最後に提案手法を評価する実験を行った。SVM、NB、BL(ベースラインモデル)、SVM+BL(SVMとベースラインモデルの組み合わせ)、SVM+NB (提案手法) の5つの手法の比較を行った。提案手法であるSVM+NBと、3つの単独の分類器の中で最も高いSVMを比較すると、再現率、F値、適用率においてはSVM+NBはSVMを上回るが、精度は劣る。特に再現率や適用率の向上が大きい。これは、SVMによる分類器が語義タグ付きコーパスにおける高頻度語のみを対象としているのに対し、SVM+NBは低頻度語に対しても語義を出力しているからである。すなわち、語義曖昧性解消が行われる単語が増加するので再現率や適用率が向上したと考えられる。一方、SVM+NBはSVM+BLと大きな差は無かった。頻度20以下の単語だけを対象に両者を比較すると、頻度が小さい単語ほどSVM+NBはSVM+BLを大きく上回った。したがって、提案手法は低頻度語の語義曖昧性解消に有効であるということが明らかになった。

目 次

第 1 章　はじめに **1**
1.1　研究の背景と目的 1
1.2　本論文の構成 2

第 2 章　関連研究 **3**
2.1　語義曖昧性解消を行う機械学習アルゴリズム 3
2.2　決定リストを用いた語義曖昧性解消に関する研究 4
2.3　辞書語義立てにおける語義曖昧性解消 5
2.4　Bootstrap による教師無し学習での語義曖昧性解消 6
2.5　検索エンジン AltaVista と WordNet を用いた語義曖昧性解消 7

第 3 章　低頻度語のための語義曖昧性解消モデルの学習 **8**
3.1　モデルの概要 8
3.1.1　モデル 10
3.1.2　素性 11
3.1.3　パラメタ推定 12
3.2　辞書定義文からの上位概念の抽出 13
3.2.1　上位概念抽出パターン 13

第 4 章　教師あり学習アルゴリズムによる分類器との混合 **19**
4.1　SVM による分類器 19
4.2　混合モデル 21

第 5 章　評価実験 **22**
5.1　辞書定義文からの上位概念 22
5.2　語義曖昧性解消実験 25
5.2.1　実験の手順 25
5.2.2　実験結果と考察 25

第 6 章　おわりに **29**

付 録 A 上位概念の抽出パターン **33**
A.1 名詞の抽出パターン 33
A.2 動詞の抽出パターン 35
A.3 形容詞の抽出パターン 38
A.4 接尾語 39
A.5 その他 40

図 目 次

3.1 コーパスに存在する文 9
3.2 コーパスに存在する文 9
3.3 人または動物を上位概念とする語義の例 9

表 目 次

5.1 上位概念の抽出 22
5.2 上位概念を抽出できた単語数 23
5.3 同じ上位概念が重複して抽出された単語数 23
5.4 実験結果 26
5.5 閾値毎における実験結果 27
5.6 低頻度語を対象にした実験結果 27

第1章　はじめに

1.1　研究の背景と目的

文章中における単語には複数の語義がある。その中から正しい語義を選択するタスクを語義曖昧性解消と呼び、自然言語処理における重要なタスクの１つである。語義曖昧性解消の応用としては、機械翻訳、自動要約など様々な応用が考えられる。例えば、「生涯をかける」と「服をタンスにかける」の２つの文章があり、我々は普通にどちらも「かける」の意味がそれぞれ違うことを理解できる。これと同じように、機械に「かける」の意味がどちらであるかを判別させる処理が語義曖昧性解消である。本研究は、文章中における単語に対して正しい語義を自動的に選択する分類器を作成することを目的とする。

語義曖昧性解消に関する過去の研究として、語義タグ付きコーパスを使った教師あり機械学習による手法が盛んに行われている。これらの手法は、単語周辺の文脈を手がかりにして語義曖昧性解消を行う分類器を学習する。語義タグ付きコーパスとは、あらかじめ正解タグをコーパスと呼ばれる電子テキストに付与していたデータである。しかし、語義タグ付きコーパス作成にはコストや時間がかかるため、大量のデータを用意することが難しい。またコーパス中に出現回数が少ない単語は分類器を学習できないというデータの過疎性の問題がある。

また本研究では、人間の文章理解を支援する読解支援システムでの使用を前提とした語義曖昧性解消のための分類器を作成する。読解支援システムでの使用が前提であることから、より多くの単語を扱えること、すなわち再現率と適用率の向上が必要である。そのため、教師あり学習による手法をそのまま用いることは適切ではない。

この問題を解決するために、本研究では辞書定義文をコーパス以外の知識源として使用する。まず、辞書定義文中の最後の表記に着目して上位概念を取り出す。次に、上位概念と周辺語の共起関係を反映した分類器を学習して語義曖昧性解消を行う。例として、犬という単語について考える。犬には２つの意味があり、それぞれの辞書定義文が「犬という動物」、「スパイという役割の人」であったとする。それぞれ辞書定義文中の文末に着目すると上位概念として「動物」と「人」が取り出せる。一般に、上位概念は複数の単語で共有される。例えば動物を上位概念とする単語は「象」「猫」「豚」と言ったように様々な単語がある。従って上位概念のコーパスにおける出現頻度は語義そのものよりも高いため、上位概念と周辺語の共起関係を手がかりとした信頼性の高い分類器を学習可能である。

一方、教師あり機械学習は低頻度語には使うことが難しいが高頻度語については高い精度で語義曖昧性解消を行うことができる。従って、本研究では、上位概念を用いる分類器

と教師あり機械学習による分類器を併用して、高精度かつ適用範囲の広いシステムを作成することを目的としている。

1.2 本論文の構成

本論文の構成は以下の通りである。2 章では、語義曖昧解消全般の関連研究について述べる。3 章では、低頻度語における辞書を使用した分類器の作成について述べる。4 章では、教師あり学習アルゴリズムによる分類器と、第 3 章で述べた分類器との混合について述べる。5 章では、システムの評価実験結果を行い、結果の考察を行う。6 章では結論と今後の課題を述べる。

第2章　関連研究

本章では語義曖昧性解消の先行研究を紹介することにする。語義曖昧性解消と関連あるタスクとしては、自然言語理解、機械翻訳、情報検索、ハイパーテキストナビゲーション、構文解析、サブカテゴリパターンの獲得、選択制限、スピーチ合成、スペル訂正、自動要約がある。語義曖昧性解消の応用例は、先ほど述べた様に様々な応用例が存在する。現在、語義曖昧性解消の先行研究は様々なものが存在し、その中で本研究との違いを述べることにする。

2.1　語義曖昧性解消を行う機械学習アルゴリズム

語義曖昧性解消 (WSD) を行う先行研究は多い。Genard らの論文 [2] では、教師あり学習というあらかじめ正解データを用いた学習による語義曖昧性解消を行う４つのアルゴリズムを評価している。特に、異る分野のコーパスを使用して学習された WSD システムを別の分野のコーパスに適用したときの評価を行っている。また、教師なし学習では、正解付きデータを用いないため精度、再現率とも良好ではない。一方、教師あり学習ではデータスパーネスの問題があるが、それを解決する手法がいくつか提案されているため、この研究では教師あり学習による手法の研究を行いたいと著者は述べている。

Gebard らは、４つの機械学習アルゴリズムを使用し実験を行った。

- Naive Bayes

$$argmax_i P(C_i| \cap v_j) \approx argmax_i P(C_i) \prod_j P(v_j|C_i) \tag{2.1}$$

  ここでは統計的学習を行い、結果を近似している。ここで C_i は語義を示し、v_j は素性を示している。

- Exemplar-based Algotrithm
  WSD を行いたい単語において、テスト文とコーパス中の類似度を計算し似ている k 個の文を取出している。k 個の文についている語義の多数決を取り、一番多かったものをその単語の語義としている。

- Snow Architecture

ニューラルネットワークによる手法を用いて、各クラスのノードから学習を行っている。

- LazyBoosting
  弱分類器を組み合わせて学習を行っている。

DSO コーパスは 121 種類の名詞と 70 種類の動詞が 192800 個含まれる語義タグ付きコーパスである。DSO コーパスは２つの異る文章を含んでおり、Wall Street Journal と Brown Corpus からなる。ここでは Wall Street Journal(WSJ) と Brown Corpus(BC) をそれぞれコーパス A、コーパス B とおき実験を行っている。コーパス A すなわち Wall Street Journal を訓練データとしコーパス B すなわち Brown Corpus をテストデータとして実験を行っている。筆者は先ほど述べた４つのアルゴリズムを用いて実験した結果、LazyBoosting が最も良い結果を出していると述べている。

本研究との違いについて述べると、教師あり学習において語義タグつきコーパスを使用し、コーパスにない単語は語義曖昧性解消ができない点である。本研究では語義曖昧性解消の頑健性を向上させることを目的としている。語義曖昧性解消のできる単語の種類を辞書から抽出した上位概念を利用する手法により増加させている。

2.2 決定リストを用いた語義曖昧性解消に関する研究

語義曖昧性解消に関する研究は多々あるが、この中で David Yarowsky[1] の決定リストを用いた語義曖昧性解消について述べる。ここではフランス語の語尾における単語のアクセントの有無の判別を行っている。David Yarowsky は前後に現れるアクセントの有無を決定する規則を獲得し、アクセントの出現傾向が極端に激しいものの規則を優先的に使用している。例えば cote の単語に着目し、前後の文脈からアクセントのパターンの判定を行っている。

- du côté
- côte ouest
- côté du gouvernement

以上の様に周りの文脈によってアクセントの付け方が変わる。Yarowsky はまず実験で曖昧なアクセントのパターンをあげ、アクセントが異る単語を挙げ違いを述べた。次に決定リストで使う素性として以下を用いた。

- 中心となる単語の 1 つ右に存在する単語
- 中心となる単語の 1 つ左に存在する単語
- 中心となるプラスマイナス K の範囲に存在する単語

- １つ左に位置する単語と２つ左に位置する単語の組
- １つ右に位置する単語と２つ右に位置する単語の組
- １つ左に位置する単語と２つ右に位置する単語の組

決定リストを学習した結果、精度は98%でありかなり良好である。

本研究との違いについて述べると、Genard らの研究と同様に教師あり学習において語義タグ付きコーパスが必要であることから、コーパスにない単語は語義曖昧性解消ができない問題点がある。

2.3 辞書語義立てにおける語義曖昧性解消

玉垣 [10] は２つの異る知識源を用いて再現率の向上を図っている。一つめの知識源として注釈付きコーパスである。注釈付きコーパスから機械学習を行い分類器を作成した。注釈付きコーパスとは、新聞記事など人手で様々な付加情報を付け加えたテキストデータであり、注釈付きコーパスから機械学習を行い分類器を作成した。注釈付きコーパスから機械学習を行うアルゴリズムとして、Support Vector Machine(SVM) を用いた。学習を行う素性として以下のものを用意し、様々な素性に対して学習を行い、最もマッチした素性を実験的に求めている。

- 多義語の前後 n 語に含まれる自立語の基本型を抜き出す。n を可変にして、最適な文脈の大きさを調査した。
- 多義語の直前、直後にある m 語の品詞情報と表記を抜き出す。m を可変にして、最適な m の大きさを調査した。多義語の前後 n 語以内に現れる自立語の意味クラスを抜き出す。意味クラスはシソーラス ID を用いた。意味クラスを用いる場合は、以下の 2 つの点で最適化を試みている。
- 一つは分類語彙表の桁数に関する最適化である。分類語彙表の ID を上位３桁から７桁まで変化させた。
- もう一つは、一つの単語が複数の意味クラスをもつ場合の処理に関する最適化である。複数の ID が存在する場合は展開して素性に加える場合と単独の ID のみを加える場合を考慮した。

またコーパス以外の知識源として岩波国語辞書を用いた分類器を組み合わせた。すなわちコーパスから学習をして作成した分類器の他に、岩波国語辞典に記述されている情報を使用して 2 種類の分類器を作成した。2 種類の分類器として、一つは用例分類器でありもう一つは文法情報分類器としている。用例を用いた分類器は、入力文と語釈文の用例の類似度を計算し、最も類似度の高い用例を持つ語義を選択している。また文法情報を用いた分

類器は、候補となる全ての語義について、入力文がその語義の文法情報を満たすかどうかを調べている。そして、文法情報を満たす語義があれば、これを正しい語義として出力する。また複数の語義が文法情報を満たすときには、これを正しい語義として出力する。

玉垣の研究と本研究の比較を行う。玉垣はコーパスに存在しない単語についても辞書を用いた分類器を用いて WSD を行う手法を提案している。これは本研究と同様である。しかし、玉垣は用例と文法情報を使うが、本研究では辞書の定義文に含まれる上位概念を使用している。辞書定義文を使用する点が玉垣との研究とは異る点である。

2.4 Bootstrap による教師無し学習での語義曖昧性解消

Yarowsky[4] は、注釈が存在しないコーパスから教師無し学習アルゴリズムによる手法により語義曖昧性解消を行っている。Yarowsky が提案した手法として、一つは文書につき意味は一つという原則と文脈につき意味は一つという原則に従って反復 Bootstrap アルゴリズムを用いている。また以下２つの点においては、最初は文脈につき意味は一つという原則について説明しており、2 つ目は文書に意味は一つという原則について説明している。

- 文脈中に存在する単語のうち意味はひとつ
- 文書中に存在する単語のうち意味はひとつ

教師無し学習アルゴリズムとして、タグ付けを行っていない複合語 7538 個の曖昧性解消を行っている。この論文では目標語が plant となる巨大なタグ無しコーパスを用意している。plant の前後には２つの主な意味のために使用されており、life が文脈の前後に存在する場合は意味を A とし、manufacuturing が存在する方を B としている。意味 A と意味 B のトレーニングデータをそれぞれセットして、決定リストで学習を行っている。決定リストにおける学習は前後 1 から 10 以内の単語において行っている。それぞれ微量のトレーニングセットにより、A か B かのどちらかの意味かを付け加えて、訓練データ量を増やす。また各 A と各 B における意味では文書毎にラベル付を行い A か B のトレーニングデータの集合を増やしている。次に反復学習を行って訓練データを成長させていき、分類できるものを収集し分類できないものを避けている。最終的に決定リストにより割り振られた意味を示している。またここでは、plant が living の意味を示しているなら A とし、factory の意味を示しているなら B の意味を指し示すようにしている。テストを行った結果は 96%を越える精度を出している。

Yarowsky と本研究との違いについて述べると、Yarowsky は、決定リストを用いた Bootstrap による教師無し学習手法による語義曖昧性解消に関する研究を行っているが、本研究では、Support Vector Machine を使用し高頻度語を対象に学習を行いまた Naive Bayes による手法を用いて低頻度を対象に学習を行い、２つの手法を組み合わせて学習を行うことにより、再現率、精度を向上させている。

2.5 検索エンジンAltaVistaとWordNetを用いた語義曖昧性解消

Micheal[3] はテキスト内に存在する名詞、動詞、副詞、形容詞を WordNet を用いて語義曖昧性解消を行っている。サーチエンジンのインターネットのヒット数を用いている。またペアとなる単語の意味的距離を WordNet を用いて評価している。文脈に存在する単語と語義曖昧性解消を行いたい単語をペアとして曖昧性を解消している。良い精度が得られる理由として、WordNet を使用し全ての単語に対して意味のランク付を行っていることが挙げられる。ここではアルゴリズム 1 とアルゴリズム 2 について説明する。アルゴリズム 1 では動詞と名詞のペアとして “investigate report” を例に挙げ説明する。動詞 investigate は文脈に存在する単語で一つの意味を持ち、report には複数の意味が存在する。そして Altavista を用いて各ペアで最も上位である候補をランク付して求めている。また各単語のペアの組みあわせを m 個用意し、各 m 個の組み合わせを AND 検索や OR 検索を用いて頻度が高いペアからランク付を行っている。次にアルゴリズム 2 について述べる。ここでは WordNet を用いて、動詞の辞書定義文中の名詞と、名詞の下位概念に相当する名詞において、共通の単語を数えることで語義曖昧性解消を行っている。アルゴリズム 1 を適用してからアルゴリズム 2 を適用する理由として、アルゴリズム 2 だけで語義曖昧性解消を行うと計算量が多すぎるため、アルゴリズム 1 を用いて意味の候補の絞りこみを行い計算量を減少させている。

Micheal らの手法と本研究の違いについて述べると、彼らは検索エンジン Altavista とシソーラスである WordNet を用いて語義曖昧性解消を行う手法を提案しているが、本研究では、EDR 概念辞書を使用し、辞書定義文から上位語を取出す手法を用いて語義曖昧性解消を行っている。また名詞、動詞、形容詞、副詞以外にも更に接尾語やサ変名詞などに対して語義曖昧性解消を行っている所が特徴である。

第3章　低頻度語のための語義曖昧性解消モデルの学習

本章では、辞書を用いた語義曖昧性解消モデルについて述べることにする。教師あり機械学習では、正解データすなわち語義タグつきデータを必要とする。そのためコーパスに現れない単語や出現頻度の低い単語には、語義を判定するモデルを学習しにくいという問題がある。そのためコーパスに適用してもあまり結果が芳しくない語義を学習するためには辞書定義文から語義の上位概念を抽出し、これを予測する確率モデルを学習することで低頻度語の語義の曖昧性解消を解消する手法を示す。

3.1　モデルの概要

本研究では、語義の定義として EDR 概念辞書 [5] による語義を用いる。EDR 概念辞書に記載されている「犬」の語義を以下に挙げる。６桁の数字は概念 ID と呼ばれる語義を識別するものである。

3bdc67 犬という 動物

3ce9f1 スパイという役割の 人

例文として、犬にえさをあげるという文の「犬」の語義を決める場合を考える。
(1) 犬 にえさをあげる
3bdc67 「動物」
3ce9f1 「人」

この文において「犬」という単語と「えさ」という単語の共起関係に着目すると、3bdc67 の意味の方が正しい語義である。

一般に、語義曖昧性を解消するモデルを正解付きデータから学習することは可能である。しかし、「犬」が一度もコーパス中に現れない場合は、「犬」の語義が 3bdc67 か 3ce9f1 を決定するモデルを学習することは不可能である。そのため本研究では辞書定義文に着目する。辞書定義文における末尾に着目すると、3bdc67 の上位概念は「動物」であり 3ce9f1 の上位概念は「人」である。この２つの上位概念により、判定を行う場合以下のことが

コーパス上に存在する文
猫にえさをあげる 0e7328 「動物」 ウサギにえさをあげる 0e475e 「動物」 亀にえさをあげる 3be548 「動物」

図 3.1: コーパスに存在する文

コーパス上に存在する文
あいつはアーチストだ 3b01cb 「人」 あいつはアイアンマンだ 3c4c2d 「人」 あいつは合方だ 0e2679 「人」

図 3.2: コーパスに存在する文

考えられる。 まず図 3.1 のような文がコーパスに存在するとする。またコーパスに正しい語義が付与されていると仮定する。「猫」の語義は 0e7378 であり、「ウサギ」の語義は 0e475e であり、「亀」の語義は 3be548 であるとする。これらの語義の辞書定義文を図 3.3 に示す。図 3.3 に示すように、語義 0e7328、0e475e、3be548 の上位概念は全て下線部の通り「動物」である。従って、コーパス上から「動物」という上位概念と「エサ」という単語がよく共起することが学習できる。そのため例文 (1) の「犬」の上位概念は、「人」ではなく「動物」であると判定でき、「犬」の語義は 3bdc97 であるということがわかる。また同様に、以下の文における語義を決める場合を考える。

猫 0e7328 猫という動物 ウサギ 0e475e ウサギという動物 亀 3be548 亀という動物
アーチスト 3d01c6 美術家という職業の人 アイアンマン 3c4c2d たくましい体と強い意志をもった人 合方 0e2679 三味線で歌の伴奏をする人

図 3.3: 人または動物を上位概念とする語義の例

(2) あいつは警察の犬だ
3bdc67「動物」
3ce9f1 「人」

図 3.2 のような文がコーパスに存在するとする。また、文中の単語に正しい語義が付与されていると仮定する。「アーチスト」の語義は 3d01c6 であり、「アイアンマン」の語義は 3c4c2d であり、「合方」の語義は 0e2679 であるとする。図 3.3 に示すように、語義 3d01c6、3c4c2d、0e2679 の上位概念は全て下線部の通り「人」である。コーパスでは「人」という上位概念と「あいつは」という単語がよく共起することが学習できる。例文 (2)「犬」の上位概念は「動物」ではなく「人」であると判断でき、語義は 3ce9f1 であると判断できる。

このように、辞書定義文から語義の上位概念を抽出し、上位概念と周辺語との共起性を学習することにより、訓練コーパスに出現しないものでも単語の語義を判定できる。なぜなら EDR 概念辞書には、「動物」や「人」を上位概念とする語義は複数存在し、これらの語義が訓練コーパスに存在するからである。また、上位概念は複数の語義で共有されるため、上位概念のコーパスにおける出現頻度は語義そのものよりも出現頻度に比べて高くなる。従って出現頻度が低く信頼性が低いモデルを学習する単語においても、上位概念を利用することにより訓練データの量を増やすことができる。

3.1.1 モデル

このページでは、辞書定義文から抽出された上位概念を用いた語義曖昧性解消モデルに付いて述べる。まず、ある単語 w の語義を決定するために、以下のような確率モデルを適用する。

$$P(s, c|F) \tag{3.1}$$

式 (3.1) において s は w の語義を示し、c は辞書定義文から抽出された s の上位概念である。また F は w を含む入力文から得られる素性集合であり、w の周辺語などの要素となる。本研究で用いた素性の具体的な内容について 3.1.2 項で述べる。次に式 (3.1) を以下のように近似する。

$$P(s, c|F) = P(s|c, F)P(c|F) \simeq P(s|c)P(c|F) \tag{3.2}$$

ここでは (3.2) の第 1 項 $P(s \mid c, F)$ を $P(s \mid c)$ として近似している。$P(s \mid c, F)$ は入力文から得られる素性集合 F と上位概念 c から語義 s を予測するモデルであり、F から s を予測するという点で Naive Bayes モデルによる語義曖昧性解消モデルとほぼ同じである。しかし、低頻度語については語義 s の出現頻度が低いため、統計的に信頼できるモデルが学習できないと考えられる。そのため、語義 s は語義の上位概念 c のみに依存するとみなして、$P(s \mid c)$ のように近似する。もう一方、式 (3.2) の第 2 項 $P(c \mid F)$ は素性集合 F から語義の上位概念 c を予測するモデルである。3.1 節で述べた通り、語義の上位概念は複数の単

語で共有されることから、語義 s よりもコーパスにおける出現頻度は高いため、$P(c \mid F)$ は低頻度語の語義曖昧性解消を行う場合でも十分学習可能である。

次にベイズの定理を用いて以下のような変形を行う。

$$P(s|c)P(c|F) = \frac{P(s)P(c|s)}{P(c)}\frac{P(c)P(F|c)}{P(F)} \tag{3.3}$$

$$= \frac{P(s)P(F|c)}{P(F)} \tag{3.4}$$

$$\simeq \frac{P(s)\prod_{f_i \in F} P(f_i|C)}{P(F)} \tag{3.5}$$

(3.3) から (3.4) の変形では $P(c \mid s) = 1$ とした。これは語義 s の辞書定義文から抽出される上位概念 c は常に一意に決まるためである。また (3.4) から (3.5) の変形において、F 中の各素性 f_i の出現は互いに独立であると仮定して近似している。

本研究では、式 (3.5) の確率を最大にする語義 s を選択することによって語義曖昧性解消を行う。ここでは、全ての語義について F は同じであり、P(F) の計算は省略可能である。

$$argmax_{s \in S_w} \frac{P(s)\prod_{f_i \in F} P(f_i|c)}{P(F)} \tag{3.6}$$

$$= argmax_{s \in S_w} P(s) \prod_{f_i \in F} P(f_i|c) \tag{3.7}$$

式 (3.7) の S_w は辞書に登録されている w の語義の集合である。直感的に言えば、式 (3.7) の第 1 項 P(s) は語義の出現頻度を学習するモデルであり、第 2 項 $\prod P(f_i \mid c)$ は語義の上位概念 c と素性 f_i の共起性を学習するモデルである。

3.1.2 素性

(3.6) のモデルに用いる素性 f_i として以下のものを用いた。いずれも語義曖昧性解消においてよく用いる素性である。なおここでいう w は語義曖昧性解消を行う単語を示している。例として「太郎の犬が吠えた」という文で、w を犬としたときの具体的な素性を示す。

- w の直前または直後の単語
  直前の単語:の
  直後の単語:が

- w の表記
  例文として、「太郎の犬が吠えてうるさい」における、「吠えて」を対象語とした時の例を挙げる。
  w:吠えて

- w の直前または直後の品詞
  直前の品詞:助詞
  直後の品詞:助詞

- 同一文中にある自立語の基本形
  自立語の表記:太郎
  自立語の表記:吠える

- w に係る格と格要素の組 (w が用言のとき)
  ここでは例文として「太郎が犬小屋を建てた」という文を示し、w を「建てた」とする。
  格と格要素の組:犬小屋/を
  格と格要素の組:太郎/が

- w の格と係り先用言の組 (w が格要素のとき)
  格と係り先用言の組:が/吠える

- 係り先文節の主辞 (w が文節の主辞のとき)
  係り先文節の主辞:吠える

- 係り元文節の主辞 (w が文節の主辞のとき)
  対象語が文節の主辞と同一であるとき、その文節の係り元文節の主辞の基本形を取出す。
  係り元文節の主辞:太郎

- 同一文節の主辞 (w が文節の主辞ではないとき)
  ここでは例文として「警察犬が吠えた」という文を示し、w を「警察」とする。
  同一文節の主辞:犬

3.1.3 パラメタ推定

式 (3.7) に示すように、本研究で推定するべき確率モデルは P(s) と $P(f_i \mid c)$ である。まず、P(s) は加算スムージングで推定したものを式 (3.8) とする。

$$P(s) = \frac{O(s) + \alpha}{\sum_s O(s) + \alpha V} \tag{3.8}$$

O(s) は語義 s の出現頻度、α は全ての事象に足すべき頻度、V は辞書中の語義の総数を示している。ここでは $\alpha = 0.5$ とした。一方、$P(f_i \mid c)$ は線形補完法で推定した。すなわち式 (3.9) のように素性の出現頻度確率 $P(f_i)$ との混合モデルとして推定する。両者の重みづけ β は 0.5 とする。$P_{MLE}(f_i \mid c)$ は式 (3.10) の最尤推定によって推定を行った。一方 $P(f_i)$ は式 (3.11) で推定した。T は訓練データの総数であり、分子と分母にそれぞれ 1 と 2 を加えているのはスムージングを行うためである。また $O(f_i)$ は素性の頻度を示し、$O(f_i, c)$ は素性と上位概念の共起関係の頻度を示している。

$$P(f_i|c) = \beta P_{MLE}(f_i|c) + (1 - \beta)P(f_i) \tag{3.9}$$

$$P_{MLE}(f_i|c) = \frac{O(f_i, c)}{\sum_{fi} O(f_i, c)} \tag{3.10}$$

$$P(f_i) = \frac{O(f_i) + 1}{T + 2} \tag{3.11}$$

モデルの学習を行うために、EDR コーパス [5] を使用している。またコーパス中の文の形態素解析と文節の係り受け解析を行うため、形態素解析ツール CHASEN[8] と係り受け解析器 Cabocha[9] を使用し、各素性を取出した。各素性を取出した後語義の頻度を $O(s)$、素性の頻度 $O(f_i)$、素性の頻度と上位概念の共起頻度 $O(f_i, c)$ をカウントし、モデルを学習した。

3.2 辞書定義文からの上位概念の抽出

3.2.1 上位概念抽出パターン

この項では、辞書定義文から上位概念を抽出する手法について述べる。基本的には辞書定義文の末尾にある単語をその語義の上位概念とみなす。また、取出すべき単語は品詞及び表記のパターンマッチにより決める。以下は抽出パターンの例である。

```
s(「N」の意を表わす語)
<*><名詞-*><*> <*><記号-括弧閉><*>* <の><助詞-連体化><*> <意><名詞-一般><*> <
を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <語><名詞-一般><*>
b:1
```

パターンについて説明すると 1 行目はパターンの識別子を示している。 2 行目は定義文にマッチする単語のパターンであり、定義文の末尾に対してマッチさせる。パターンの要

素は＜基本形＞＜品詞（第１レベル）＞＜品詞（第２レベル）＞で指定を行っている。品詞体系として茶筌の第一レベルと第二レベルの品詞を用いた。なお＜ ＞の中に存在する “*” は任意の基本形や品詞にマッチすることを表している。一方、＜ ＞の外の “*” は 0 個以上の単語にマッチすることであり、“+” は 1 個以上の単語にマッチすることを示している。[ ] は他の抽出パターンを使用することを表している。[ ] を用いた抽出パターンの例を以下に挙げる。

```
n(N のこと)
[n(複合名詞),n] <の><助詞-連体化><*> <こと><名詞-非自立-一般><*>
o:1
```

[ ] の中には使用する抽出パターンの識別子を書く。[ ] 内に複数の抽出パターンがある場合には、抽出パターンを左から順番に用いる。例えば、上記の抽出パターンで、[n(複合名詞),n] は他の抽出パターン「n(複合名詞)」を適用し、マッチングに失敗したらさらに抽出パターン「n」を適用することを表す。３行目はマッチングに成功した時に取出す上位概念である。「アルファベット:数字」は上位概念をどのように抽出するかを示すための表記である。「数字」は２行目の抽出パターンにおける要素の左からの位置を表す。一方、「アルファベット」は以下のような意味を持つ。h は表記を示し、o は [ ] で示された他のパターンによって抽出された上位概念を示し、b は基本形を示す。bb は複合語でマッチしたとき、先頭の語の基本形だけを取り出す。また oh は o と同じだが、最後の単語は基本形ではなく表記を取り出す。また、パターンにはあらかじめ順番があり、その順番に従って抽出パターンを適用させる。以上で、上位概念を抽出するパターンを説明した。以下は、単語「令」の辞書定義文に対して s(「N」の意を表わす語) のパターンを適用した。
下線部はパターンの 2 行目にマッチした単語を表わし、→ の右は抽出された上位概念を表わす。

ex:令 「公布された <u>命令」の意を表わす語</u> → 命令

一方、以下は単語「因縁」の辞書定義文に対して、n(N のこと) のパターンを適用した例である。

ex: 因縁 欠点をさがしだして言う <u>悪口のこと</u> → 悪口

ここでは上位概念抽出パターンを 64 種類人手で作成した。品詞別に抽出パターンの一部を取り上げる。残りは付録 A に示すことにする。

名詞

名詞の辞書定義文は多くの場合名詞で終わるので、文末にある名詞を上位概念として取出す。ただし複数名詞の場合は最後の名詞のみを上位概念として抽出する。これをおこな

うのがパターン n である。

n
<*><名詞-*><*>
b:1

ex:電圧 アーク電圧という、溶接アークにかかる 電圧 → 電圧
　またパターン n(複合名詞) の場合、最後の名詞だけでは上位概念が曖昧としてふさわしくないため複合名詞として取出している。例えば、以下の辞書定義文の末尾における「物」だけを取出すと意味が広いため、複合名詞「廃棄物」を抽出している。

n(複合名詞)
名詞
<*><名詞-*><*>* <物|法><名詞-*><*>
h:1 b:2

ex: 産業廃棄物 産業活動において生じる 廃棄物 → 廃棄物
名詞の辞書定義文が体言ではなく動詞で終わる場合がある。特にサ変名詞の辞書定義文に良くみられる。名詞の辞書定義文が動詞で終わっているとき、動詞のパターンを使って上位概念を取出す場合がある。そして動詞の上位概念に「こと」をつけて補い、名詞として取り扱う。例として、「相乗」の辞書定義文の上位概念抽出方法を以下に示す。

v
<*><動詞-*><*>+
bb:1

ex:相乗 一つの乗り物に一緒に 乗る → 乗る+こと

一方、n(V たばかりであること) は、辞書定義文の末尾に動詞と「こと」がある場合「こと」を含めて上位概念を取出すパターンである。
n(V たばかりであること)
[o(動詞基本パタン 3)] <た|だ><助動詞><特殊・タ> <ばかり><助詞-副助詞><*> <だ><助動詞><特殊・ダ> <ある><助動詞><*> <こと><名詞-*><*>
o:1 + こと

ex: 掃立 掃除を 終えたばかりであること → 終える+こと

動詞

動詞の辞書定義文は多くの場合動詞で終わるので、文末に現れる動詞を上位概念として取出す。ただし、複合動詞の場合は先頭の動詞のみ取出す。これを行うのがパターン v である。

```
v
<*><動詞-*><*>+
bb:1
```

ex: 洗いきる 汚れを完全に 取り除く → 取る

また、動詞の後に形式名詞が続く場合は、形式名詞を取り除いて上位概念を取出す。
これを行うのがパターン v(V+形式名詞) である。

```
v(V+形式名詞)
[o(動詞基本パタン 3)] <こと|さま|もの|物|程度><名詞-*><*>
o:1
```

ex: 網打する 投網を打って魚を とること → とる

形容詞

形容詞の辞書定義文は「ーするさま」のように用言の後に「さま」が続くものが多い。このとき、用言が形容詞ならその形容詞を抽出パターン a(A さま) で取出す。また、用言が形容詞以外なら、抽出パターンとして a(V+さま) を使用し、「用言+さま」を上位概念として取出す。

```
a(A さま)
<*><形容詞-*><*> <さま><名詞-*><*>
b:1
```

ex:細かい （形が）非常に 小さい さま → 小さい

```
a(V+さま)
[o(動詞基本パタン 3)] <さま><名詞-非自立-一般><*>
o:1 + さま
```

ex:苦しい 心痛に痛みを 感じるさま → 感じる+さま

接尾語

接尾語の上位概念は、基本的には名詞と同じパターンを使用して抽出している。ただし、接尾語の辞書定義文に特有の表現がいくつかあったため、これらと別のパターンを用意した。例として、s(N の単位) と s(N を表わす語) を示す。

カラットには 2 つの語義がある。それらの辞書定義文を示す。

カラット カラットという宝石の重さの単位
カラット カラットという金の純度の単位

最後の名詞を取出すと両方とも「単位」になり、2 つの語義が何の単位か判別できない。その前に現れる名詞も含めた方が上位概念として適していると考えられる。そのため抽出パターン s(N の単位) を作成した。

```
s(N の単位)
<*><名詞-*><*> <の><助詞-連体化><*> <単位><名詞-一般><*>
b:1 + 単位
```

ex: カラット カラットという宝石の 重さの単位 → 重さの単位
ex: カラット カラットという金の 純度の単位 → 純度の単位

また「を表す語」という定義文に対応するためにパターン s(N を表わす語) を作成した。

```
s(N を表わす語)
<*><名詞-*><*> <を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <語><名詞-一般><*>
b:1
```

ex: イニング 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 → 回数

形容動詞

形容動詞は名詞と同じパターンを用いて上位概念を抽出した。

```
n(V た+形式名詞)
[o(動詞基本パタン 3)] <た|だ><助動詞><特殊・タ> <こと|さま|もの|物|程度><名詞-*><*>
o:1 + 完了 + b:
```

ex: 赤塗 赤く 塗ったもの → 塗った+もの

副詞

副詞に関しては、辞書定義文のパターンが形容詞に類似しているので、形容詞のパターンを使用した。例を以下に挙げる。

抽出パターン
a(V+さま)
[o(動詞基本パタン3)] <さま><名詞-非自立-一般><*>
o:1 + さま

ex:愛らしい 小さくて愛らしいさま 愛らしいさま → 愛らしい+さま

サ変名詞

サ変名詞は、文中では名詞としても動詞としても使われる可能性があるが、名詞として使われているときは、名詞の抽出パターンを使用する。また動詞として使用されている時は、動詞の抽出パターンを使用する。
名詞としての抽出パターンを用いた例文を挙げる。以下の文における「研修」は名詞として使われている。従って、名詞の上位概念が取出されている。

今日は社内 研修 があった。

以下は「研修」の辞書定義文から上位概念を抽出する様子を示している。

v(S をする)
動詞
<*><名詞-サ変接続><*> <を><助詞-格助詞-一般><*> <する><*><*>
h:1 b:3

ex:研修する 必要な知識を身につけるため特別な 勉強をする → 勉強する+こと

この場合、動詞の抽出パターン v(S をする) で上位概念を取出しているので、抽出した上位概念の末尾に「こと」をつけることにより名詞として上位概念を取出している。
また動詞としての抽出パターンを用いた例文を挙げる。以下の文における「研修」は動詞として使われている。従って、動詞の上位概念が取出されている。
社員を 研修 する。

ex:研修する 必要な知識を身につけるため特別な 勉強をする → 勉強する

第4章　教師あり学習アルゴリズムによる分類器との混合

4.1　SVMによる分類器

本節では語義タグ付きコーパスから分類器を機械学習する手法を提案する。教師あり学習アルゴリズムとして Support Vector Machine(SVM) を使用した。学習に用いた素性は以下の通りになる。またこれらの素性を得るため、形態素解析器として茶筌 [8] を、文節の係り受け解析器として Cabocha[9] を用いた。例として「太郎の犬が吠えた」と言う文で w を犬としたときの具体的な素性を示す。

- w の直前と直後の単語
  対象語及びその周辺にある語の表記を示している。また対象語からプラス 2 語、マイナス 2 語の表記も示している
  マイナス 2 語の単語:太郎
  直前の単語:の
  直後の単語:が
  プラス 2 語の単語:吠えた

- w の表記
  例文として「太郎の犬が吠えてうるさい」における、「吠えて」を対象語としたときの例を挙げる。
  w:吠えて

- w の直前または直後の品詞
  w を中心に直語、マイナス 2 語の周辺にある語や直後の単語やプラス 2 語の品詞を示している。品詞は茶筌の品詞体系を使用した。
  マイナス 2 語の単語の品詞:名詞
  直前の単語の品詞:助詞
  直後の単語の品詞:助詞
  プラス 2 語の単語の品詞:動詞

- w の直前または直後に現れる２つの単語の組
  対象語の直前、直後にある２つの語の表記の組を示している。
  直前の２つの単語の組:太郎/の
  直後の２つの単語の組:が/吠えた
  直前、直後にある単語の組:の/が

- w の直前または直後に現れる２つの品詞の組
  対象語の直前、直後にある２つの語の品詞の組を示している。
  直前の２つの単語の品詞の組:名詞/助詞
  直後の２つの単語の品詞の組:助詞/動詞
  直前、直後にある単語の品詞の組:助詞/助詞

- 同一文中にある自立語の基本形
  自立語の表記:太郎
  自立語の表記:吠える

- 係り先文節の主辞 (w が文節の主辞のとき)
  対象語が文節の主辞と同一であるとき、その文節の係り先文節の主辞の基本形を取出す。
  係り先文節の主辞:吠える

- 係り元文節の主辞 (w が文節の主辞のとき)
  対象語が文節の主辞と同一であるとき、その文節の係り元文節の主辞の基本形を取出す。
  係り元文節の主辞:太郎

- 同一文節の主辞 (w が文節の主辞ではないとき)
  対象語が文節の主辞でないとき、その文節の主辞の基本形。
  ここでは例文として「警察犬が吠えた」という文を示し、w を「警察」とする。
  同一文節の主辞:犬

- w に係る格と各要素の組
  ここでは例文として「太郎が犬小屋を建てた」という文を示し、w を「建てた」とする。
  格と格要素の組:犬小屋/を

格と格要素の組:太郎/が

- w の格と係り先用言の組 (w が格要素のとき)
  w が名詞で、かつある用言の格要素となっているとき、その格と動詞の基本形の組。
  格と用言の組:が/吠える

なお、SVM に用いる素性と Naive Bayes モデルで用いた素性 (3.1.2) 項は若干異り、SVM に用いる素性は若干多い。SVM に追加している素性は

- w の直前と直後の単語におけるプラス 2 語、マイナス 2 語の単語

- w の直前または直後の品詞におけるプラス 2 語、マイナス 2 語の品詞

- w の直前または直後に現れる 2 つの単語の組

- w の直前または直後に現れる 2 つの品詞の組
  である。
  なぜなら、実験において、(3.1.2) で述べた素性のみを使うよりも、これらの素性を追加した方が精度がよかったからである。

SVM の学習に LIBSVM[1] を用いた。ν-SVM [7] によって学習を行い、カーネルは線形カーネル、$\nu = 0.0001$ とした。SVM は 2 値分類器であるのに対して、本研究における語義曖昧性解消問題は多値分類問題である。そこで、pairwise 法を用いて SVM を多値問題に適用した。

4.2 混合モデル

本研究は、高頻度語のための SVM による分類器と低頻度語のための上位概念を用いた分類器の 2 つを組み合わせる。組み合わせる手法として以下を試みた。語義を決めたい単語の訓練データにおける出現頻度が閾値以上なら SVM による分類器を、それ以外は辞書定義文の上位概念を用いて判別を行う。実験ではこの閾値を 20 とした。SVM を用いた機械学習では高頻度、すなわち良く出現する語義に関しては、結果は良好であるが、低頻度の単語に対しては精度が良くない。そのため低頻度の単語においては辞書定義文を用いた Naive Bays モデルを使用し精度及び再現率の向上を試みている。

[1] http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/

第5章　評価実験

5.1　辞書定義文からの上位概念

3.2.1 項で述べた抽出パターンを用いて、EDR 日本語単語辞書ならびに EDR 概念辞書 [5] にある概念説明 (辞書定義文) から上位概念を抽出した。辞書定義文の形態素解析には茶筌 [8] を用いた。結果を表 5.1 に示す。

表 5.1 の「語義数」の行は EDR 日本語単語辞書に含まれる語義の総数、「上位概念抽出語義数」の行は定義文から上位概念を取出すことのできた語義の数を示している。また、括弧内の数値は辞書にある語義のうち上位概念を取出すことのできた語義の割合を示している。名詞や動詞についてはほとんどの語義の上位概念を取出すことができたが、その他の品詞については 7 割から 9 割程度の語義に対してしか上位概念を抽出することができなかった。上位概念の抽出に失敗する主な原因は抽出パターン不足であり、抽出パターンを追加することによってある程度対処できると思われる。

一方、1 つの上位概念は複数の語義の定義文から取出される可能性がある。そこで「上位概念異り数」と「平均語義数」の値を調べた。「上位概念異り数」は語義の定義文から取出された上位概念の異り数を示している。「平均語義数」の行は上位概念抽出語義数を上位概念異り数で割った値であり、同じ上位概念が抽出される語義数の平均を示している。例えば、名詞の場合、平均 7 個の語義の定義文から同じ上位概念が抽出されている。上位概念ができるだけ多くの語義で共有されていればいるほど、すなわち平均語義数が大きければ大きいほど、低頻度語に対して訓練データの量を増やす効果が大きいと考えられ

表 5.1: 上位概念の抽出

品詞	語義数	上位概念抽出語義数 (割合)	上位概念異り数	平均語義数
名詞	170746	169668(0.9937)	24210	7.0082
動詞	31024	30351(0.9783)	8763	3.4635
形容詞	1529	1236(0.7828)	645	1.9163
接尾語	1575	1381(0.8768)	684	2.0190
形容動詞	5197	4578(0.9155)	2268	2.0979
副詞	3055	2204(0.7214)	937	2.3522
全ての品詞	195066	191266(0.9783)	35389	5.4047

表 5.2: 上位概念を抽出できた単語数

全て抽出	一部抽出	抽出失敗
87449(0.9455)	3844(0.0415)	1013(0.0109)

表 5.3: 同じ上位概念が重複して抽出された単語数

重複なし	一部重複	全て重複
74319(0.8111)	8479(0.0925)	8821(0.0962)

る。大ざっぱに見積もれば、辞書定義文から抽出された上位概念を用いることにより、訓練データの量を約 5 倍に増やす効果があると言える。

本研究で辞書定義文から上位概念を抽出する理由は、語義曖昧性解消を行うモデルを学習するためである。そのため語義曖昧性解消にどれだけ有効かという観点から抽出された上位概念を評価する。ある単語があった時、その全ての語義の辞書定義文から上位概念を抽出することができない場合、語義曖昧性解消に関しては有効ではない。その例として「あたかも」の単語の辞書定義文を以下に挙げる。

0e3178 その時ちょうど

0f19cd あたかも

ここでは２つの語義のいずれからも上位概念が抽出できなかったから語義曖昧性解消には適切ではない。そのため単語単位での上位概念の抽出の成功割合を示したのが表 5.2 である。表 5.2 の「全て抽出」は、単語が複数の語義を持つとき、全ての語義から上位概念を抽出できた単語の数である。その例として「２階建て」における語義の辞書定義文を以下に示す。

102032 二階建ての建物 ⇒ 建物

102033 二階のある建物の構造 ⇒ 構造

それぞれ上位概念として、「建物」と「構造」の 2 つが抽出されている。以上のように全ての辞書定義文から上位概念が抽出できている。次に「一部抽出」は一部の語義からのみ上位概念を抽出できた単語数である。その例として「あげる」における辞書定義文を以下に示す。

3c2ce4 へりくだる意を添える語

3cf6fc 仕事や物事を終える ⇒ 終える

3ce6db 物事や行為を成し遂げること ⇒ 成し遂げる

ここでは上位概念として、「終える」と「成し遂げる」が抽出できているが、3c2ce4 だけが上位概念の抽出に失敗しているため、一部抽出となっている。最後に「抽出失敗」は全ての語義について上位概念を取出すことができなかった単語数である。その例は上であげた「あたかも」である。また括弧内の数値は EDR 日本語単語辞書に含まれる多義語の総数の内、それぞれが抽出された単語の割合を示している。表 5.2 より抽出失敗は 1%程度

であるから、本研究で提案した上位概念抽出手法は適切な手法だと思われる。

一方、ある単語が複数の語義を持つ時に、その全ての語義から同じ上位概念が重複して抽出される場合は、提案手法は語義曖昧性解消に有効ではない。例として「あくた火」の辞書定義文を以下に挙げる。

0e2d0c ごみを焼く火 ⇒ 火

0e2d0d 漁夫が藻屑を焼く火 ⇒ 火

これらの辞書定義文から抽出された上位概念は「火」であり、２つの上位概念は重複していることが明らかである。このように辞書定義文から同じ上位概念が全て重複する場合、どちらの語義が正しい語義かを判別できないという問題点が生じる。そのため単語単位で、上位概念が重複して抽出される割合を示したのが表 5.3 である。表 5.3 の「重複なし」は単語が複数の語義を持ちかつ 1 つ以上の語義から上位概念を取出すことができたとき、その全ての語義に対して異る上位概念を抽出できた数を示している。その例として「２人作家」の辞書定義文を挙げる。

1f5e60 一つの作品を共同で執筆する 2 人の作家 ⇒ 作家

1f5e5f 2 人の作家が 1 作品を執筆すること ⇒ 執筆する+こと

この辞書定義文より、上位概念は重複していないことがわかる。また「一部重複」は一部の語義について同じ上位概念を抽出できた単語数を示している。その例として「８mm」の辞書定義文を挙げる。

103cf3 ハチミリという光学機械 ⇒ 機械

103cf4 八ミリ撮影機という、幅が八ミリメートルのフィルムを使用して撮影する光学機械 ⇒ 機械

103cf2 八ミリ映画という、映画の方式 ⇒ 方式

これらの辞書定義文では、上位概念として「機械」が重複して存在する一方、「方式」という別の上位概念が存在している。最後に「全て重複」は全ての語義から同じ上位概念を抽出した単語数を表わす。その例は上で挙げた「あくた火」である。また括弧内の数値は EDR 日本語単語辞書に含まれる多義語の総数の内、それぞれが抽出された単語の割合を示している。表 5.3 の「重複なし」に相当する単語は、語義毎に異る上位概念が抽出されており、これらを用いることにより語義曖昧性解消の正解率向上が期待できる。一方「一部重複」に相当する単語についても、正しいと思われる語義の数を絞り込む効果があると考えるなら、上位概念が語義曖昧性解消の正解率を引き上げる効果があると期待できる。従って、両者をあわせた約 89%の単語について、上位概念を抽出する効果がある。

また抽出した上位概念がどれだけ適切であるかを判定するため 200 個ランダムに選んで人手で判定を行った。例として、適切であると思われる上位概念を一部抜粋する。

200e11 日本人の氏名 ⇒ 氏名

0efaaa 憲兵隊という, 旧日本陸軍の軍隊 ⇒ 軍隊

0efbc4 天上から見た, 人間の住んでいる世界 ⇒ 世界

0f03fc 学校外での生活指導 ⇒ 指導

0f0863 工場に属する財産の集合 ⇒ 集合

また上位概念として、不適切な例を以下に挙げる。
3c3b82 飛行場で, 乗客の昇降口 ⇒ 口
3c2183 鋼鉄でつくってあること ⇒ ある+こと
その結果、上位概念は 200 個中、185 個が正解であった。
結論として、実験は間違いが 15 個程度であり全体の 92.5%が上位概念を正しく抽出できていることから結果は良好であると言える。

5.2 語義曖昧性解消実験

5.2.1 実験の手順

実験にはEDR コーパス [5] を使用した。EDR コーパスは約 20 万文からなるコーパスであり、各単語に EDR 概念辞書の概念 ID が付与された語義タグ付きコーパスである。EDR コーパスのうち、20000 文をテストデータとし、161322 文を訓練データとした。テストデータに含まれる評価単語数は 91986 である。実験では５つの手法を用いて語義曖昧性解消を行った。

- SVM(Support Vector Machine)
  4.1 節で述べたような SVM を使用した手法。主に高頻度で出現する単語の語義を判別する分類器。具体的には頻度 5 以上の単語について SVM を学習した。

- NB(Naive Bayes)
  3 章で述べたような辞書定義文から取り出した上位概念を用いた Naive Bayes モデルによる分類器。

- BL(Baseline)
  最も出現頻度が高い語義を優先的に選択する分類器。但し、最頻出語義が複数ある場合にはその全てを答えとして選択している

- SVM+NB
  4.2 節で述べたように高頻度 (5 以上) で出現する単語には SVM を使用し、低頻度語 (4 以下) の語義を判別するために NB を用いる分類器。

- SVM+BL
  高頻度 (5 以上) に出現する単語には SVM を使用し、頻度が 4 以下の単語については BL で語義の判別を行う分類器。

5.2.2 実験結果と考察

テストデータに対する語義曖昧性解消の再現率、精度、F 値、適用率を表 5.4 に示した。なお再現率、精度、F 値、適用率のそれぞれの詳細を示す。

表 5.4: 実験結果

	SVM	NB	BL	SVM+NB	SVM+BL
再現率	70.03	63.79	61.80	71.83	71.56
精度	72.11	66.50	61.38	71.78	71.60
F 値	71.06	65.12	61.69	71.81	71.58
適用率	97.12	95.78	99.76	99.98	99.77

- 精度 (P) = システムが出力して正解であった語義数/システムが出力した語義数
- 再現率 (R) = システムが正解を出力した単語数/テストデータに含まれる単語数
- F 値 = 2PR/P+R
- 適用率 = システムが正解、不正解に関わらず語義を出力した単語数/テストデータに含まれる単語数

なお表 5.4 の数値は 100%を最大とする数値を示している。考察として、SVM と BL を比較すると再現率、精度、F 値については SVM が上回っている。NB と BL を比較すると、再現率、精度、F 値については BL よりも NB が高いことが明らかである。また SVM、NB、BL の中では、SVM が適用率を除く再現率、精度、F 値が 3 つの中で一番良いことがわかる。なお適用率が若干良くないのは、SVM が高頻度語を対象にしたものであるからである。

提案手法である SVM+NB と、3 つの単独の分類器の中で最も F 値が高い SVM を比較すると、精度以外の評価値において SVM+NB は SVM を上回る。特に再現率や適用率の向上が大きい。これは、SVM による分類器が語義タグつきコーパスにおける高頻度語のみを対象としているのに対し、SVM+NB は低頻度語に対しても語義を出力しているからである。従って語義曖昧性解消が行われる単語が増加するので再現率や適用率が向上したと考えられる。一方 SVM と比べて SVM+NB の精度は劣る。しかし再現率がそれ以上に向上した。F 値も改善されていることがわかる。

教師あり学習に基づく頑強性を向上させる方法として、BL(ベースライン) の向上との併用も考えられる。そこで、SVM による分類器とベースラインモデルとの混合モデルを作成し、提案手法との比較を行った。SVM+NB の組み合わせと SVM+BL の組合わせのモデルを比較すると SVM+NB の方が精度、再現率、F 値を上回っていることが表 5.4 より明らかである。しかし、その差は大きいとは言えない。

また混合モデルにおける閾値として、頻度 5 以上を SVM の分類器で評価し、頻度 4 以下を Naive Bayes モデルでの分類器で評価したが、閾値として最適かどうかが問題であるため、閾値を調節しながら各頻度ごとに実験を行い表 5.5 に示した。表 5.5 は閾値を調節し、各閾値毎に SVM+NB と SVM+BL での混合モデルにおける F 値を示している。表 5.5 の F 値を比較すると、SVM+NB との混合モデルの方が、SVM+BL との混合モデル

表 5.5: 閾値毎における実験結果

頻度	5	6	7	8	9	10
F 値 (SVM+NB)	71.80	71.80	71.77	71.76	71.76	71.71
F 値 (SVM+BL)	71.58	71.51	71.44	71.38	71.38	71.22

頻度	11	12	13	14	15	16
F 値 (SVM+NB)	71.68	71.67	71.65	71.66	71.65	71.60
F 値 (SVM+BL)	71.17	71.12	71.07	71.05	71.00	70.93

頻度	17	18	19	20
F 値 (SVM+NB)	71.59	71.58	71.56	71.53
F 値 (SVM+BL)	70.88	70.86	70.81	70.73

表 5.6: 低頻度語を対象にした実験結果

頻度	0	1	2	3	4
単語の異り数	444	467	452	404	382
平均 F 値 (SVM+NB)	0.4911	0.6461	0.6441	0.6506	0.6538
平均 F 値 (SVM+BL)	0.2747	0.5508	0.5554	0.5536	0.6165

の値を上回っていることがわかる。ここでは閾値 5 が SVM+NB の混合モデルにおいて F 値が最も高いことから閾値としてピークであると言える。閾値 4 以下を調べていない理由として、頻度 4 以下の単語を学習しようと場合、SVM は単語毎に分類器を学習するため単語数が増大し現実的に学習が適用できる範囲ではないためである。次に頻度 5 の F 値 (SVM+BL) と頻度 18 の F 値 (SVM+NB) において両者は値が同じであることが解る。ここでは頻度 18 の SVM+NB のモデルの方が、頻度 5 の SVM+BL と比べて SVM で作成する分類器の数が少なくてすむ。具体的には、頻度 18 では単語数は 4378 個であり、頻度 5 では単語数は 8345 個である。同程度の F 値なら SVM+NB の方が SVM が学習する単語数が少なくてすむということが挙げられる。

SVM+NB と SVM+BL の差がそれほど大きくない理由として、低頻度語と高頻度語を合わせて表示したためあまり変化が見られないようになっている。そこで低頻度語に対する実験を行った。結果を表 5.6 に示した。表 5.6 の「頻度」は訓練コーパスに含まれる単語の出現頻度を示している。「単語の異り数」はテストデータ内に出現する単語の内、訓練コーパスにおける頻度が表 5.6 の第 1 列の頻度であるような単語の異り数を示している。「平均 F 値」は訓練コーパスにおける頻度が表 5.5 の第 1 列の頻度であるような単語について、それぞれの単語の F 値の平均を示している。SVM+NB の分類器を使用した場合は「平均 F 値 (SVM+NB)」であり、SVM+BL の分類器を使用した場合は「平均 F 値 (SVM+BL)」である。

単語の異り数は頻度が増加していくごとに減少傾向にある一方、表 5.6 の SVM+NB と SVM+BL の平均 F 値を比較すると、SVM+NB の方が平均 F 値が高いことが明らかである。また、頻度の低い単語ほど平均 F 値の差が大きいことがわかる。例として頻度 0 における SVM+NB と SVM+BL の平均 F 値を比較すると、頻度 0 から 4 までの中で最も差が大きいことが明らかである。この実験により、SVM+NB と SVM+BL を比べるとやはり SVM+NB の方が良いといえる。

第6章　おわりに

本論文では、より多くの単語を扱うことのできる語義曖昧性解消手法を提案した。2種類の分類器を作成し、これらを組み合わせることを試みた。2種類の分類器のうち、高頻度語はSVM分類器を用いて実験を行った。また低頻度語はEDR概念辞書定義文を用いて分類器を実装した。

最後に、今後の課題を４つ挙げる。

1. 辞書定義文を用いた語義曖昧性解消
   今回の実験では、EDR概念辞書中の上位概念を取出すため人手で作成した抽出パターンを用いた。抽出に失敗した例が若干あったが９割近くの単語の上位概念を取出すことができた。今後は上位概念の抽出パターンの改良を行う必要がある。
2. SVM+NBの組み合わせ手法
   本論文では、SVM+NBとの組み合わせを行い実験を行った。結果はベースラインモデルよりも精度並びに再現率の向上が見られた。今回SVMは閾値を20とし高頻度語を対象に実験を行った。また低頻度語にはNaive Bayesを用いて実験を行った。今回の実験では閾値を20と定め実験を行ったが、閾値をこれよりも低くし実験を行う方法や他の組み合わせ手法を検討するべきである。
3. 教師無し学習との比較
   本研究と教師無し学習を比較すると、本研究は辞書の上位概念を使用することにより再現率と精度を向上させている。今回は低頻度語すなわち、単語の出現回数が低いものに関してもベースラインによる手法よりも精度と再現率を向上させることができた。しかし、語義タグ付きコーパスを必要としない教師無し学習での手法との実験的な比較を行ったわけではない。SVMによる教師あり学習とは別に教師無し学習に関しても検討したい。
4. 異る分類器の組み合わせ
   玉垣[10]は、岩波国語辞書の用例、文法情報などを使用した分類器とSVMによる分類器を組み合わせることにより再現率の向上を試みた。本研究では、EDR概念辞書の定義文を用い上位概念の抽出を行い抽出された上位概念を用いた分類器を作成している。玉垣などの岩波国語辞書を用いた分類器と本研究における分類器を組み合わせることによる手法も検討したい。

謝辞

本研究を進めるにあたり，終始熱心な御指導を賜わりました白井清昭助教授に心から感謝致します．さらに数多くの御教授を頂きました島津明教授に厚く御礼申し上げます．山田寛康助手ならびに自然言語処理学講座の皆様には，貴重な御意見、討論をして頂きました事を感謝致します.

関連図書

[1] David Yarowsky,Decision List for Lexical Ambiguity Resolution Application to Accent Restoration in Spanish and French ,ACL1994 p88-p95

[2] Escudero, G. , L. Mrquez and G. Rigau. On the Portability and Tuning of Supervised Word Sense Disambiguation Systems. In Proceedings of the Joint Sigdat Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, Hong Kong. 2000.

[3] Mihalcea, R., and Moldovan, D. A method for Word Sense Disambiguation of unrestricted text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99) (Maryland, NY, June 1999),1999.

[4] David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, Meeting of the Association for Computational Linguistics,189-196,1995

[5] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書第２版
Technical Report TR-405,1995

[6] 池原悟、宮崎正弘、白井諭、横尾昭男、中岩浩己、小倉健太郎、大山敦史林良彦、日本語語彙体系 - 全５巻 -. 岩波書店,1997

[7] Bernhard Scholkopf, Alex J. Smola, R. Williamson, and P. Barlett. New support vector algorithms. Neural Computation, Vol.12, pp1083-1121,2000.

[8] 松本裕治、北内啓、山下建雄、平野喜降、松田寛、高岡一馬、浅原正幸、茶筌 version 2.3.3 使用説明書,2003

[9] 日本語係り受け解析システム「南瓜」 マルチメディア言語学情報 [18], 月刊 言語, Vol.32, No.6, pp.74-75, June 2003.

[10] 玉垣 隆幸, 辞書の語義立てに基づく語義曖昧性解消に関する研究 修士論文 2004

付 録A　上位概念の抽出パターン

品詞別の上位概念抽出パターンの一覧を示す。
なお、抽出パターンの適用順序はここで挙げた抽出パターンの順序と一致する。

A.1　名詞の抽出パターン

名詞から上位概念を抽出するパターンの一覧を以下に示す。

n(N のこと)
[n(複合名詞),n] <の><助詞-連体化><*> <こと><名詞-非自立-一般><*>
o:1

n(V ていない+形式名詞)
[o(動詞基本パタン 2)] <て><助詞-接続助詞><*> <いる><動詞-非自立><*> <ない><助動詞><*> <こと|さま|もの|物><名詞-非自立-一般><*>
o:1 + 否定 + b:5

n(V ない+形式名詞)
[o(動詞基本パタン),a] <ない><助動詞><*> <こと|さま|もの|物><名詞-非自立-一般><*>
o:1 + 否定 + b:3

n(V た+形式名詞)
[o(動詞基本パタン 3)] <た|だ><助動詞><特殊・タ> <こと|さま|もの|物|程度><名詞-*><*>
o:1 + 完了 + b:3

n(V+形式名詞)
[o(動詞基本パタン 3)] <こと|さま|もの|物|程度><名詞-*><*>
o:1 + b:2

n(A+形式名詞)
[a] <こと|さま|もの|物|程度><名詞-*><*>
oh:1 + b:2

n(C+形式名詞)
<*><名詞-*><*> <*><助動詞><特殊・ダ> <こと|さま|もの|物><名詞-非自立-一般><*>
h:1 + コピュラ + b:3

n(V たばかりであること)
[o(動詞基本パタン3)] <た|だ><助動詞><特殊・タ> <ばかり><助詞-副助詞><*> <だ><助動詞><特殊・ダ> <ある><助動詞><*> <こと><名詞-*><*>
o:1 + こと

n(N である+形式名詞)
[o(名詞)] <だ><助動詞><特殊・ダ> <ある|ない><助動詞><*> <こと|さま|もの|物|程度><名詞-*><*>
h:1 + コピュラ + b:4

n(N の程度)
[n,n(A さ)] <の><助詞-*><*> <程度><名詞-一般><*>
o:1

n(A さ)
<*><形容詞-*><*> <さ><名詞-接尾-特殊><*>
h:1 b:2

n(V+方)
<*><動詞-自立><*> <方><名詞-接尾-特殊><*>
h:1 b:2

n(V ぐあい)
<*><動詞-自立><*> <ぐ><未知語><*> <あい><名詞-一般><*>
h:1 ぐあい

n(複合名詞)
<*><名詞-*><*>* <物|法><名詞-*><*>
h:1 b:2

n
<*><名詞-*><*>
b:1

A.2　動詞の抽出パターン

動詞から上位概念を抽出するパターンの一覧を以下に示す。

v(V ようになる)
[o(動詞基本パタン 2)] <た|だ><助動詞><特殊・タ>* <よう><名詞-非自立-助動詞語幹><*> <に><助詞-副詞化><*> <なる><動詞-自立><*>
o:1

v(V そうになる)
[o(動詞基本パタン),v(V せる),v(V れる)] <そう><名詞-接尾-助動詞語幹><*> <に><助詞-格助詞-一般><*> <なる><動詞-自立><*>
o:1

v(N しようとする)
[o(名詞)] <する><動詞-*><*> <う><助動詞><*> <と><助詞-格助詞-一般><*> <する><動詞-自立><*>
o:1 b:2

v(V ようとする)
<*><動詞-*><*> <う><助動詞><*> <と><助詞-格助詞-一般><*> <する><動詞-自立><*>
b:1

「〜したり、〜したりする」という並列のパタン。
　現状だと 2 番目の動詞しか抽出しない
v(V たりする)
[o(動詞基本パタン 2)] <たり|だり><助詞-並立助詞><*> <する><動詞-自立><*>
o:1

ex. 「消えてなくなる」 → ない + なる
v(なくなる)
<てる><動詞-非自立><*> <ない><助動詞><特殊・ナイ> <なる><動詞-自立><*>
ない + なる

ex. 「引けなくなる」 → 引く + 否定
v(V なくなる)
[o(動詞基本パタン)] <ない><助動詞><特殊・ナイ> <なる><動詞-自立><*>
o:1 + 否定

v(N になる)

[o(名詞)] <に><助詞-*><*> <なる><動詞-自立><*>
b:1 + になる

v(ADJ なる)
<*><形容詞-*><*> <なる><動詞-自立><*>
b:1 + なる

v(V ことができる)
[o(動詞基本パタン 3)] <こと><名詞-非自立-一般><*> <が><助詞-格助詞-一般><*> <できる><動詞-*><*>
o:1

v(V+形式名詞)
[o(動詞基本パタン 3)] <こと|さま|もの|物|程度><名詞-*><*>
o:1

v(V て+非自立動詞)
[o(動詞基本パタン 2)] <て><助詞-接続助詞><*> <いる|いく|おく|しまう><動詞-非自立><*>
o:1

v(V で+非自立動詞)
[v] <で><助詞-接続助詞><*> <いる|いく|おく|しまう><動詞-非自立><*>
o:1

v(V+非自立動詞)
[o(動詞基本パタン 2)] <始める|はじめる|続ける|過ぎる|すぎる|合う|かける|終わる|終える><動詞-非自立><*>
o:1

v(V れる)
[o(動詞基本パタン)] <れる|られる><動詞-接尾><*>
o:1 + 受身

v(V ようにさせる)
[o(動詞基本パタン 2)] <よう><名詞-非自立-助動詞語幹><*> <に><助詞-格助詞-一般><*> <する><動詞-*><*> <せる><動詞-接尾><*>
o:1 + 使役

v(Vせる)
[v,v(Sする),v(ADJする),v(Sをする),v(Vをする),v(Nをする),v(Nにする)] <させる><動詞-*><*>
[o(動詞基本パタン)] <せる|させる><動詞-接尾><*>
o:1 + 使役

v(ADJする)
<*><形容詞-*><*> <する><*><*>
b:1 + する

v(Sする+動詞)
[n,o(ADV)] <する><*><*> <*><動詞-*><*>
o:1 b:2

v(Sする)
[n,o(ADV)] <する><*><*>
o:1 b:2

v(V連用+する)
<*><動詞-自立><*> <する><*><*>
b:1

v(Sをする)
<*><名詞-サ変接続><*> <を><助詞-格助詞-一般><*> <する><*><*>
h:1 b:3

v(Vをする)
<*><動詞-自立><*> <を><助詞-格助詞-一般><*> <する><*><*>
b:1

v(Nをする)
[n,o(ADV)] <を><助詞-格助詞-一般><*> <する><*><*>
o:1 + をする

v(Nがある)
[o(名詞)] <が|の><助詞-格助詞-一般><*> <ある|ない><*><*>
o:1 + が b:3

v(N にする)
動詞
[o(名詞)] <に><助詞-格助詞-一般><*> <する><*><*>
o:1 + に b:3

保留
v(N+P する)
<*><名詞-*><*> <*><助詞-格助詞-一般><*> <する><*><*>
h:2 + する

v
<*><動詞-*><*>+
bb:1

A.3 形容詞の抽出パターン

形容詞から上位概念を抽出するパターンの一覧を以下に示す。
a(V+ないさま)
[o(動詞基本パタン 3)] <ない><助動詞><*> <さま><名詞-非自立-一般><*>
o:1 + 否定 + さま

a(V+さま)
[o(動詞基本パタン 3)] <さま><名詞-非自立-一般><*>
o:1 + さま

a(C+さま)
<*><名詞-*><*> <*><助動詞><特殊・ダ> <さま><名詞-非自立-一般><*>
h:1 + コピュラ + さま

a(N であるさま)
<*><名詞-*><*> <だ><助動詞><特殊・ダ> <ある|ない><助動詞><*> <さま><名詞-*><*>
h:1 + コピュラ + さま

a(ADV であるさま)
<*><副詞-*><*> <だ><助動詞><特殊・ダ> <ある|ない><助動詞><*> <さま><名詞-*><*>
h:1 + コピュラ + さま

a(A さま)
<*><形容詞-*><*> <さま><名詞-*><*>
b:1

a
<*><形容詞-*><*>
b:1

A.4 接尾語

接尾語から上位概念を抽出するパターンの一覧を以下に示す。
s(〜の一つ)
[s(A さの単位),s(N の単位),s(A さを表わす単位),s(N を表わす単位),s(N 単位)] <の><助詞-連体化><*> <一つ><名詞-一般><*>
o:1

s(A さの単位)
<*><形容詞-*><*> <さ><名詞-接尾-特殊><*> <の><助詞-連体化><*> <単位><名詞-一般><*>
h:1 b:2 + 単位

s(N の単位)
<*><名詞-*><*> <の><助詞-連体化><*> <単位><名詞-一般><*>
b:1 + 単位

s(A さを表わす単位)
<*><形容詞-*><*> <さ><名詞-接尾-特殊><*> <を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <単位><名詞-一般><*>
h:1 b:2 + 単位

s(N を表わす単位)
<*><名詞-*><*> <を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <単位><名詞-一般><*>
b:1 + 単位

s(「N」の意を表わす語)
<*><名詞-*><*> <*><記号-括弧閉><*>* <の><助詞-連体化><*> <意><名詞-一般><*> <を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <語><名詞-一般><*>

b:1

s(N を表わす語)
<*><名詞-*><*> <を><助詞-格助詞-一般><*> <表す><動詞-自立><*> <語><名詞-一般><*>
b:1

s(N 単位)
<*><名詞-*><*> <単位><名詞-一般><*>
b:1 + 単位

s(語)
<語><名詞-一般><*>
*

A.5 その他

その他における上位概念を抽出するパターンを以下に示す。

品詞のパタンマッチには使わないが、他のパタンマッチ規則の部品として使う
o(名詞)
[n,o(ADV),o(V)]
oh:1

o(ADV)
<*><副詞-*><*>
b:1

o(V)
<*><動詞-自立><*>
b:1

「動詞基本パタン 2」 + α (動詞全て)
o(動詞基本パタン 3)
[o(動詞基本パタン 2),v(V ようになる),v(V そうになる),v(N しようとする),v(V ようとする),v(V たりする),v(V なくなる),v(V ようにさせる),v(V て+非自立動詞),v(V で+非自立動詞),v(V+非自立動詞)]
o:1

「動詞基本パタン」 + 受身形、使役形、〜なる、〜ある

o(動詞基本パタン 2)
[o(動詞基本パタン),v(V せる),v(V れる),v(N になる),v(ADJ なる),v(N がある)]
o:1

o(動詞基本パタン)
[v,v(S する),v(S する+動詞),v(V 連用+する),v(ADJ する),v(S をする),v(V をする),v(N をする),v(N にする)]
o:1