分類語彙表の分類項目を識別する語義曖昧性解消

北陸先端科学技術大学院大学
先端科学技術研究科

小林 健人

平成 30 年 3 月

修 士 論 文

分類語彙表の分類項目を識別する語義曖昧性解消

1610226 小林 健人

主指導教員 白井 清昭

審査委員主査 白井 清昭
審査委員 東条 敏
飯田 弘之
池田 心

北陸先端科学技術大学院大学

先端科学技術研究科 [情報科学]

平成 30 年 2 月

概 要

語義曖昧性解消 (Word Sense Disambiguation;WSD) は，複数の語義を持つ単語が文中に出現したとき，その文脈で使われている語義を選択する問題である．この際，単語の語義は辞書やシソーラスによって定義される．日本語の著名なシソーラスの一つに分類語彙表がある．分類語彙表は，語を意味によっていくつかの分類項目に分けている．分類項目は，それに該当する意味を持つ単語を列挙することで定義されている．分類語彙表は日本語を対象とした自然言語処理に広く利用されていることから，分類語彙表の分類項目を語義としたWSD技術は利用価値が高い．しかし，このような技術はこれまでそれほど盛んに研究されていなかった．本論文は，分類語彙表の分類項目を識別する語義曖昧性解消を実現することを目的とする．教師あり機械学習に必要な正解付きの訓練データを用意することが難しいことから，本研究では教師なし機械学習の手法を探究する．

まず，提案手法のベースとなるYarowskyの手法について述べる．この手法は，サブコーパスの構築と特徴の獲得という2つのステップからなる．サブコーパスの構築では，分類項目毎に，それに属する単語を含む文をコーパスから抽出し，分類項目のサブコーパスとする．この際，対象単語の前後20単語を文とみなして抽出する．次に，分類項目毎に，その意味を表わす特徴を獲得する．ここでの特徴とは，分類項目の意味を持つ単語の文脈に出現しやすい自立語と定義する．すなわち，サブコーパス中の用例の文脈に出現する自立語を分類項目の特徴として獲得する．さらに，特徴の重みを推定する．特徴の重みとは，特徴が分類項目の意味をどの程度顕現的に表わすかを示す指標である．上記の手続きでWSDの分類モデルの学習が完了する．語義を決めたい対象語を含む文 (テスト文) が与えられたとき，対象語の周辺に出現する分類項目の全ての特徴について，その重みの和を分類項目のスコアとし，そのスコアが最大の分類項目をその対象語の語義として選択する．

さらに，Yarowskyの手法に対して，以下の3つの拡張手法を提案する．第1に，コロケーションを特徴として用いる．コロケーションは，対象語とその前後に出現する単語から構成される単語列であり，WSDに有効な特徴であることが知られている．Yarowskyの手法はコロケーションをWSDに用いていないが，本研究はこれを特徴として用いることで，特に動詞のWSDの正解率を向上させることを狙う．

第2に，サブコーパスを構築する際に単義の単語のみを用いる手法を提案する．Yarowskyの手法では，訓練の際には単語の多義性を解消しないため，分類項目の特徴として正しくないものが獲得されやすいという問題点がある．これに対し，本研究では単義の単語のみを訓練に使用することで，正しくない特徴が獲得されるのを妨げることを狙う．

第3に，訓練データを漸進的に増加させる手法を提案する．まず単義の単語のみを用いてWSDのモデルを学習する．次に，そのモデルを用いて訓練データにおける多義語の語義を推定する．語義が1つに決まり，かつその判定の信頼度が高いとき，その事例を新たに訓練データに追加する．次に，増加した訓練データを用いてWSDのモデルを再学習する．これを繰り返すことで，訓練データを漸進的に増加させる．以下，この手法を「語義

推定モデル」と呼ぶ．これに加え，上記の手続きにおいて，訓練データにおける多義語の語義を 1 つに絞るのではなく，判定スコアの低い (つまり不正解の可能性が高い) 語義を削除する手法も提案する．以下，これを「語義絞り込みモデル」と呼ぶ．

提案手法の評価実験について述べる．テストデータとして，Balanced Corpus of Contemporary Written Japanese (BCCWJ) に分類語彙表の分類項目を付与したコーパスを用いる．テストデータにおける対象語の数は，名詞が 2467，動詞が 1175，その他が 270，合計 3912 である．Yarowsky の手法の正解率は 51.9%であった．提案手法のうち，コロケーションを特徴に加えたモデルの正解率は 51.7%であった．全体の正解率は低下したが，動詞のみを対象としたときの正解率は 44.3%であり，Yarowsky の手法の 40.9%を上回った．単義の単語のみを訓練データとしたモデルの正解率は 56.7%であり，Yarowsky の手法を 4.8 ポイント上回った．訓練データを漸進的に増加させる手法のうち，「語義推定モデル」は単義の単語のみを訓練データとしたモデルと比べて，正解率が 1.8 ポイント向上し，「語義絞り込みモデル」では正解率が 1 ポイント向上した．本研究の提案手法は Yarowsky の手法よりもおおむね正解率が高いことから，提案手法の有効性が確認できた．

目 次

第1章　はじめに

1.1　研究の背景

複数の意味（語義）を持つ単語がある特定の文脈に出現したとき，その文脈で使われている単語の語義を識別する処理は語義曖昧性解消 (Word Sense Disambiguation; 以下，WSD と記す) と呼ばれている．WSD は自然言語処理の中で最も重要な基礎技術の一つである．WSD を必要とする典型的な自然言語処理応用システムとして機械翻訳が挙げられる．いま，日英機械翻訳システムにおいて,「味が染みて煮崩れる」という日本語文を英文に翻訳する場面を考える．まず，入力された日本語文を形態素解析することで,「味/が/染みて/煮崩れる」という単語列を得ることができる．機械翻訳の基本的な処理のひとつは，日本語の単語を英語の単語に翻訳することである．このとき,「染みて」という単語を英単語に訳すときに曖昧性が生じる．すなわち，内部に液体，においなどが染み込む意味の permeate と，心に刺激が染みる意味の sink deeply in mind という 2 つの対訳の候補がある．これは,「染みる」という単語が 2 つの意味を持ち，それぞれの意味によって英語の対訳が異なっているからとみなせる.「味が染みて煮崩れる」という文においては,「染みる」の意味は前者であることを推定し，すなわち語義の曖昧性を解消し，permeate と訳さなければ，自然な翻訳は生成できない．このように，WSD は自然言語処理において重要な役割を担っている．

一般に WSD では，単語の語義は辞書やシソーラスによって定義される．日本語の著名なシソーラスに分類語彙表がある．分類語彙表では，およそ 81000 の単語が 900 個の意味カテゴリに体系的に分類されている．分類語彙表では，似た意味を持つ単語を集めた意味カテゴリのことを「分類項目」と呼ぶ．分類語彙表は日本語を対象とした自然言語処理に広く利用されていることから，多義語に対して分類語彙表の分類項目を決める WSD 技術は利用価値が高い．しかしながら，分類語彙表の分類項目を推定する WSD 手法はこれまであまり研究されていない．これは，近年の WSD の研究は教師あり機械学習に基づく手法が主流であるのに対し，分類語彙表の分類項目が付与された大規模な語義付きコーパスが存在しなかったことが一因と考えられる．現代日本語書き言葉均衡コーパス (Balanced Corpus of Contemporary Written Japanese; 以下，BCCWJ と記す) に対して分類語彙表の分類項目をアノテーションする試みが進められているが [2]，研究者に広く利用できる段階には至っていない．

1.2 研究の目的

本研究は，分類語彙表の分類項目を語義の定義とした語義曖昧性解消手法を提案する．教師あり機械学習に必要となる分類語彙表の分類項目が付与された大規模なコーパスが存在しないため，本研究では教師なしの機械学習手法を探求する．また，教師あり機械学習には，正解率は比較的高いものの，語義を識別するための分類器を個々の単語毎に学習し，また分類器を学習できるのは訓練データの量が多い高頻度語に限られるため，低頻度語については語義を推定することができないという問題もある．これに対し，本研究では，分類語彙表に登録されている全ての単語の語義の曖昧性を解消する手法を探究する．

本研究の目的である分類項目を語義の定義としたWSDの具体例を下記の例文を用いて説明する．

(例) 味 が [染みて] 煮崩れる

WSDの対象単語を [染みて] とし，この単語の分類項目を推定する場合を考える．「染みて」の基本形は「染みる」である．「染みる」という単語は，分類語彙表では表1.1に示す2つの分類項目に登録されている．これは「染みる」が2つ以上の複数の意味を持つ多義語であることを意味している．

表 1.1: 「染みる」の分類語彙表における分類項目

分類番号	分類項目
2.1533	漏れ・吸入
2.3002	感動・興奮

2.1533ならびに2.3002は分類項目に与えられた識別番号である．分類語彙表ではこれを分類番号と呼ぶ．ここでの目標は，2.1533と2.3002の2つの分類項目のうち，正しい意味を選択することである．2.1533の分類項目は「液体が浸透する」という意味に，2.3002の分類項目は「心が揺れ動かされる (例. 心に染みる)」という意味に相当する．したがって，この例文の場合は，2.1533の分類項目を選択する方が正解となる．

前述したように，教師あり機械学習では，単語毎にWSDの分類器を学習する．上記の例文では，「味」「染みる」「煮崩れる」という単語毎に分類器を学習する．これは，単語によって語義，すなわち分類クラスが異なるためである．訓練データにおける出現回数が少ない単語については，その単語の語義を識別する分類器は学習できない．本研究では，高頻度語，低頻度に関わらず，文中の全ての単語 (自立語) の語義を推定できるモデルを学習する．

1.3 本論文の構成

本論文の構成を以下に述べる．第2章では，分類語彙表を紹介するとともに，関連研究について述べ，本研究との違いについて論じる．第3章では，本論文の提案手法について

述べる．提案手法のベースとなる Yarowsky の手法の詳細を説明し，WSD の正解率を向上させるために行った改良を詳述する．第 4 章では，提案手法の評価実験について述べ，実験結果を考察する．最後に第 5 章では，本論文の成果を総括し，今後の課題について述べる．

第2章　関連研究

本章では，WSD の関連研究について述べる．2.1 節では，本研究で語義の定義とする分類語彙表について説明する．2.2 節では，教師あり機械学習を用いた WSD に関する先行研究を紹介する．2.3 節では，教師なし機械学習を用いた WSD に関する先行研究を紹介する．最後に 2.4 節では，本研究と先行研究の違いについて議論する．

2.1　分類語彙表

一般的な WSD の問題設定では，単語の語義はあらかじめ定義されている．この際，1 章で述べたように，辞書やシソーラスによって語義を定義することが多い．本研究では，分類語彙表によって語義を定義する．分類語彙表とは，国立国語研究所資料集として刊行された日本語シソーラスである．1964 年に初版 [3] が発行され，自然言語処理など幅広い分野で活用されている．本論文では，2004 年 9 月発行された「分類語彙表 -増補改訂版-」[4] を使用した．

分類語彙表は，日本語の単語を意味によって分類・整理したシソーラスである．分類語彙表では分類項目が定義されている．分類項目とは，似た意味を持つ単語の集合 (カテゴリ) である．分類項目には 5 桁の分類番号が付与される．さらに，分類番号が近い分類項目は互いに似ている意味を持つという性質を持つ．分類項目の例を表 2.1 に示す.

表 2.1: 分類語彙表における分類項目の例

分類項目	分類番号	単語
漏れ・吸入	2.1533	染みる, 吸う, にじむ, 注ぎ足す, ...
感動・興奮	2.3002	染みる, 驚く, 怒る, 感動する, ...

分類番号が 2.1533 の分類項目は「漏れ・吸入」であり，液体や気体の移動を伴う動作に関する意味を持つ単語として「染みる」「吸う」「にじむ」「注ぎ足す」などの単語が登録されている．また，分類番号が 2.3002 の分類項目は「感動・興奮」であり，人間の感情変化を伴う単語として「染みる」「驚く」「怒る」「感動する」などの単語が登録されている．

分類語彙表における分類項目の数は 895，分類項目に登録されている単語ののべ数は 81,320 である．ひとつの分類項目に登録されている単語数の平均はおよそ 90 である．

一般に，一つの単語が複数の分類項目に分類されることがある．例えば,「染みる」という単語は表 2.1 に示したように 2.1533 と 2.3002 の両方の分類項目に分類される．この

とき，その単語は複数の語義を持つ多義語であるとみなすことができる．一方，単語が1つの分類項目にしか属さないときは，その単語は1つの語義しか持たない単義語とみなせる．

2.2 教師あり機械学習に基づく語義曖昧性解消

Yuan らは，Long Short-Term Memory(LSTM) に基づく手法，ならびにそれを拡張した半教師あり機械学習による語義曖昧性解消手法を提案した [1]．機械学習において，文を素性ベクトルで表現する際によく用いられるのは Bag-of-words(BOW) モデルである．このモデルでは，文中に出現する単語を素性ベクトルの次元とし，文の意味を表現する．このとき，各単語の出現は独立に扱われ，隣接する単語の情報や単語間の統語的関係は失なわれる．この問題に対し，Yuan らは，文の単語列を時系列とみなして，これを LSTM の入力とすることにより，テキスト内の単語の並びを考慮して WSD を行うモデルを学習している．これにより，単語の並びや単語間の統語的関係を考慮した WSD モデルを学習することが可能である．例えば，統語的関係にある名詞と動詞の組が特定の語義の文脈に出現するといった傾向を学習できる．さらに，訓練データの量が不足するという問題を解消するために，半教師ありのラベル伝搬法を提案した．半教師あり機械学習は，ラベル付き学習データとラベル無し学習データの2つを用いて行う．Yuan らは，ウェブ上から大量のテキストを取得した．取得したウェブ上のテキストには語義が付与されていないため，これはラベル無し学習データである．ラベル付きデータとラベル無しデータに出現する用例から1つのグラフを構築し，ラベル付きデータの用例に付与されている語義をラベル無しデータの用例に伝播する．これをラベル伝播法と呼んでいる．この操作により，ウェブ上のテキストに語義が付与され，これも教師データとして利用する．結果として，LSTM による WSD モデルの学習とラベル伝搬を用いた半教師あり機械学習を組み合わせることにより，既存の研究と比較して 7.9 ポイントの精度向上を達成できたと報告している．

一般に，訓練データとテストデータのドメインが異なるとき，両者のドメインが同じときよりも，訓練データで学習したモデルをテストデータに適用したときの正解率が低下することが知られている．Komiya らは，日本語の WSD を行う際に，複数のドメインのコーパスから WSD のモデルを個別に学習し，対象の単語が属するテキストのドメインを判断し，それと同じドメインの訓練データから学習したモデルを選択する手法を提案した [5]．ここでのドメインとは，新聞，白書，ウェブなどといったテキストのジャンルである．テキストのジャンルによって語義の使われ方が異なるため，WSD においても，訓練データとテストデータのドメインが異なると正解率が低下する．この問題に対し，テスト文のドメインを自動判定し，同じドメインの訓練データを常に用いるという解決策を示した．評価実験の結果，異なるドメインの訓練データから学習したモデルを選択的に用いる手法は，全てのドメインのコーパスを合わせた訓練データから学習されたモデルを常に使う手法よりも，WSD の精度が向上することを確認した．

2.3　教師なし機械学習に基づく語義曖昧性解消

Yarowsky は，ロジェのシソーラスを語義の定義とした著名な WSD 手法を提案した [9]．ロジェのシソーラスは，1024 個のカテゴリに対して，そのカテゴリに該当する単語を列挙することで語の意味を分類した英語のシソーラスである．この手法では，ロジェのシソーラスのカテゴリに対し，そのカテゴリに属する単語の周辺に特徴的に出現する単語，およびその単語とカテゴリの関連度を学習する．テスト文の単語の語義を決める際には，その単語の文脈に出現する単語とカテゴリの関連度を基に，最も適切なカテゴリを選択する．ロジェのシソーラスは，単語を列挙することで語義を定義している点は日本語のシソーラスである分類語彙表と共通している．本論文は，Yarowsky の手法を分類語彙表に適用し，分類語彙表の分類項目を識別する WSD を実現する．Yarowsky の手法の詳細は 3.1 節で述べる．

鈴木らは，分類語彙表の分類項目を識別する WSD の手法を提案した [8]．テスト文に含まれる多義語を A とするとき，A が属する分類項目における A 以外の語を類義語 B と定義する．A の文脈，および訓練コーパスにおける類義語 B の文脈を単語の分散表現を基にベクトルで表現する．このとき，コーパスにおける類義語 B の出現は分類項目 (語義) の出現とみなす．k-NN 法によって，すなわち A の文脈と近い k 個の類義語を検索し，それらの類義語が属する分類項目の多数決によってテスト文における A の分類項目 (語義) を決定する．さらに，訓練コーパスにおける多義語の語義を上記の方法で推定し，単語を分類項目に置き換え，そのコーパスから分類項目の分散表現を word2vec で推定し，これを k-NN 法における事例間の類似度計算に利用する手法も提案している．評価実験の結果，WSD の正解率は 56.3%〜59.6%となった．

2.4　本研究の特色

近年の WSD に関する研究は，2.1 節で紹介したような教師あり機械学習に基づく手法が主流である．しかし，既に述べたように，分類語彙表の分類項目がタグ付けされた大規模な語義タグ付きコーパスは存在しない．そのため，本研究では教師なし機械学習に基づく手法を探究する．

Yarowsky の手法 [9] は，ロジェのシソーラスのカテゴリを語義の定義とした WSD 手法であるが，ロジェのシソーラスと分類語彙表の類似性から，これを分類語彙表の分類項目を語義とした WSD に適用することが可能である．ただし，ロジェのシソーラスは英語のシソーラスであるのに対し，分類語彙表は日本語のシソーラスである．Yarowsky の手法を日本語のシソーラスに適用したとき，どの程度の正解率が得られるかは自明ではない．本研究では，Yarowsky の手法を分類語彙表に適用し，その性能を実験的に評価する．さらに，Yawowsky の手法の問題点を考察し，それを解決するために Yarowsky の手法を拡張する手法を提案する．その拡張手法のひとつは，Yuan らの手法 [1] のように，語義付きデータを自動的に構築するものである．

鈴木らの手法 [8] も，本研究と同じく，分類語彙表の分類項目を識別する WSD 手法である．本研究のアプローチは鈴木らの手法とは異なるが，大規模なコーパスから分類項目の分散表現の学習を繰り返す彼らの手法と比べて，少ない計算時間で WSD モデルを構築できるという特徴がある．

第3章　提案手法

本章では，分類語彙表の分類項目を識別する語義曖昧性解消手法を提案する．3.1 節では，ベースの手法となる Yarowsky の手法を説明する．3.2 節では，本論文で提案する Yarowsky の手法の拡張手法について説明する．最後に，3.3 節では，分類語彙表の辞書引きについて提案システムの実装の詳細を述べる．

3.1　Yarowsky の手法

本項では Yarowsky の手法 [9] を説明する．この手法は英語のロジェのシソーラスを語義の定義として用いる．一方，本論文では，日本語のシソーラスである分類語義表の分類項目を語義の定義とする．既に述べたように，ロジェのシソーラスと分類語彙表は，同じ意味を持つ単語のグループ (カテゴリもしくは分類項目) を定義している点で共通しているため，Yarowsky の手法を分類語彙表の分類項目の WSD に適用することが可能である．以降の説明では，分類語彙表の分類項目を例に Yarowsky の手法の詳細を述べる．

語義の分類モデルの学習は，図 3.1 に示すように,「サブコーパスの作成」と「分類項目の特徴の獲得＆重みの計算」の 2 つのステップから構成される．

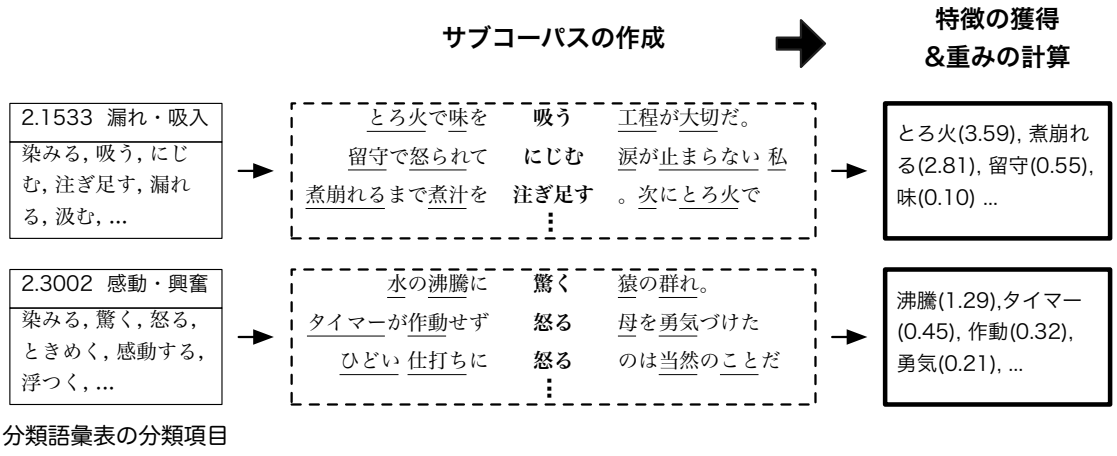

図 3.1: Yarowsky の手法 [9] の概要

3.1.1　サブコーパスの作成

分類語彙表の分類項目毎に，その分類項目に登録されている単語を含む用例を学習用コーパスから検索・収集し，サブコーパスを作成する．図 3.1 では，2.1533 の分類項目に登録されている「吸う」「にじむ」「注ぎ足す」などを含む例文が収集されている．一方，2.3002 の分類項目については,「驚く」「怒る」などを含む例文が収集されている．分類項目に登録されている単語を中心に，その前後 20 単語を例文として収集した．このサブコーパスは，分類項目の意味を持つ単語が出現する文脈を集めたものとみなせる．

本研究では，学習用コーパスとして現代日本語書き言葉均衡コーパス (BCCWJ)[6] を用いた．BCCWJ は，国立国語研究所によって作成された，現代日本語の書き言葉のテキストを集めたコーパスである．出版物や図書などの書籍全般，雑誌，新聞，白書，ブログ，ネット掲示板，教科書，国会会議録など様々な種類のテキストから構成されている．表 3.1 に学習に用いた BCCWJ のファイルの一覧と，それぞれの単語数を示す．本論文で使用した BCCWJ は 13 種類のドメイン (文書のジャンル)，36 個のファイルで構成されている．

分類語彙表における全 895 個のうち 838 個の分類項目に対し，BCCWJ から合計 2497 万件の用例を獲得した．1 つの分類項目に対するサブコーパスの用例数の平均は 29,800 であった．一方，57 個の分類項目については用例をひとつも獲得できなかった．

表 3.1: BCCWJ のファイルの一覧

ドメイン	ファイル名	単語数
書籍 (出版)	PB1	6909546
	PB2	6980662
	PB3	6883203
	PB4	7021168
	PB5	5897002
雑誌	PM	5390926
新聞	PN	1615076
書籍 (図書館)	LBe	1473538
	LBr	2065348
	LBo	2210679
	LBa	834579
	LBg	1624001
	LBd	1278036
	LBc	1084515
	LBl	2272786
	LBi	1927365
	LBf	1537172
	LBm	2090728
	LBp	1921401
	LBs	2165114
	LBb	1034586
	LBk	2018516
	LBj	2043037
	LBh	1805712
	LBn	2154028
	LBt	2108468
	LBq	2154481
白書	OW	5693403
教科書	OT	1125388
広報紙	OP	4697015
ベストセラー	OB	4434404
Yahoo!知恵袋	OC	12066093
Yahoo!ブログ	OY	13067279
韻文	OV	233457
法律	OL	1206120
国会会議録	OM	5599178

3.1.2 特徴の獲得

3.1.1 項で作成されたサブコーパスより，分類項目毎に，その意味を表す特徴を獲得する．ここでの特徴とは，分類項目の意味を持つ単語の文脈に出現しやすい自立語と定義される．すなわち，サブコーパス中の用例の文脈に出現する自立語を分類項目の特徴として獲得する．サブコーパス作成の際，分類項目の登録単語の前後 20 単語を用例として獲得した．したがって，特徴を抽出する対象となる文脈の長さは前後 20 単語である．このとき，文の境界は無視して前後 20 単語の文脈を定義している．つまり，別の文に出現する自立語でも，分類項目の登録語の前後 20 単語の範囲に出現していれば，その分類項目の特徴として獲得する．すなわち，ここでの特徴とは，同じ文に出現する単語ではなく，分類項目の登録単語が含まれるテキスト全体に出現する単語と定義している．

さらに特徴の重みを推定する．特徴の重みとは，特徴が分類項目の意味をどの程度顕現的に表すかを示す指標である．分類項目 c における特徴 f の重み $w(c,f)$ は式 (3.1) のように定義される．

$$w(c,f) = \log \frac{Pr(f|c)}{Pr(f)} \tag{3.1}$$

$Pr(f|c)$ は分類項目 c のサブコーパスにおける f の出現確率を表し，式 (3.2) のように定義される．一方，$Pr(f)$ はコーパス全体における f の出現確率を表し，式 (3.3) のように定義される．

$$Pr(f|c) = \frac{n_f^c}{\sum_{f \in D_c} n_f^c} \tag{3.2}$$

$$Pr(f) = \frac{n_f^{all}}{\sum_{f \in D_{all}} n_f^{all}} \tag{3.3}$$

これらの式において，D_c は分類項目 c のサブコーパス，D_{all} はコーパス全体，n_f^c は c のサブコーパスにおける特徴 f の出現頻度，n_f^{all} はコーパス全体における特徴 f の出現頻度を表す．

特徴および重みを求める際には，特徴の多義性を解消しないことに注意していただきたい．すなわち，ある単語が複数の分類項目 (例えば c_1 と c_2) に登録されているとき，それを含む用例は c_1 と c_2 のサブコーパスのいずれにも含まれ，また用例の文脈に出現する自立語は c_1，c_2 の両方の特徴として獲得される．一方，単語がある文脈に出現するときは 1 つの意味で使われるため，このやり方は誤った特徴を取得する可能性を排除できない．図 3.2 に例を示す．いま,「これまで会う機会がない」という例文が訓練コーパスにあったと仮定する．この中の「会う」という単語は以下の 3 つの分類項目に登録されている．

- 2.1550 合体・出会い・集合
  「2 つものが合わさる、集まる」といった意味を表わす.

- 2.3310 人生・禍福
  「運命などにめぐりあう」といった意味を表わす.

- 2.3520 応接・送迎
  「人と会う，人に対応する」といった意味を表わす．

この例文の中では,「会う」の正しい分類項目は 2.3520 である．しかし，Yarowsky の手法では，多義語の曖昧性を解消せず，この例文を 2.1550，2.3310，2.3520 の全ての分類項目のサブコーパスに追加する．したがって，正解の分類項目 2.3520 については，この例文から正しい特徴が獲得できるが，不正解の分類項目 2.1550 と 2.3310 については，正しくない特徴が獲得されることになる．

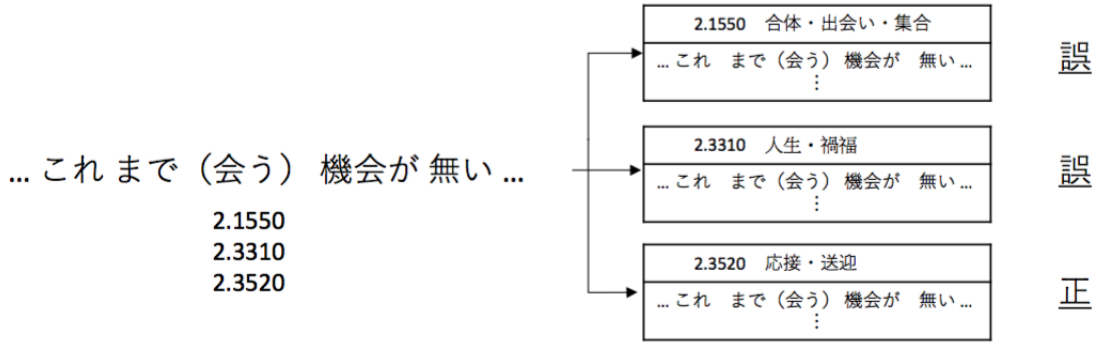

図 3.2: 誤った特徴が獲得される例

しかし，Yarowsky は，大量のコーパスから特徴を取得すれば，訓練データ中の語の多義性を解消しないことの悪影響は軽減されると主張している．また，正しくない特徴が取得される影響を軽減するために，サブコーパスにおいて一つの単語の用例を k 回取得したときは，その用例の文脈に出現する特徴は $1/k$ 回出現したとみなして $Pr(f|c)$ を推定している．図 3.3 は,「1.3850 技術・設備・修理」という分類項目に対するサブコーパスである．1.3850 に登録されている単語のひとつは「整備」だが，サブコーパスでは「整備」を中心とする用例が 3 つ獲得されている．このとき，これらの用例の文脈に出現されている特徴の出現頻度を 1/3 として $Pr(f|c)$ を計算する．

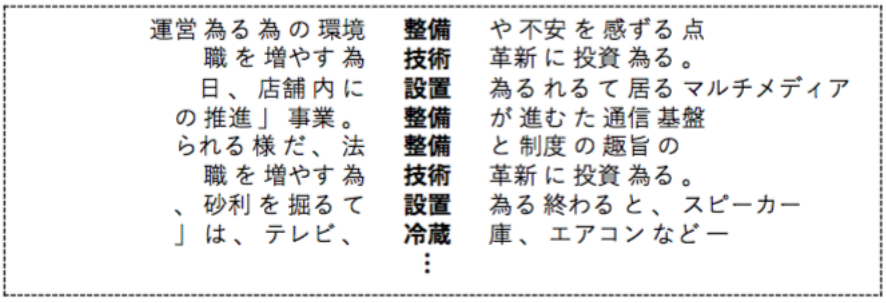

図 3.3: 分類項目「1.3850 技術・設備・修理」のサブコーパス (一部)

3.1.3 語義の推定

Yarowsky の手法では，分類項目に対する特徴とその重みのリストが語義の分類モデルとなる．本項では，語義のモデルを学習後，これを未知の文に適用して多義語の語義を解消する手法について説明する．語義を決めたい対象語を含む文 (テスト文)s が与えられたとき，式 (3.4) で定義される分類項目のスコア $score(c)$ を求め，それが一番大きい分類項目を選択する．式 (3.4) における f はテスト文の文脈に出現する特徴を表す．

$$score(c) = \sum_{f\ in\ s} w(c, f) \tag{3.4}$$

以下，語義を決定する処理の具体例を説明する．今，以下の文における [染み] の語義を決める場合を考える．
(例)タイマー で 沸騰 を 保つ 事は とろ火 で味 が [染み] て 煮崩れる こと が なく 留守 でも 安心 。

上記の例文における [染み] の基本形は「染みる」である.「染みる」は分類項目として図 3.4 に示した 2.1533 と 2.3002 の分類項目を持つ多義語である．2.1533 と 2.3002 のそれぞれについて，図 3.4 に示すように「染みる」の文脈中に出現する下線の特徴の重みを求める．次に式 (3.4) の通りに特徴の重みの和を求め，$score(c_i)$ を計算する．この場合，c_1=2.1533 のスコアが c_2=2.3003 のスコアよりも大きいため，2.1533 を [染み] の分類項目 (語義) と決定する．

2.1533　漏れ・吸入　c_1		2.3002　感動・興奮　c_2	
特徴　f	重み $w(c_1, f)$	特徴　f	重み $w(c_2, f)$
とろ火	3.59	沸騰	1.29
煮崩れる	2.81	タイマー	0.45
留守	0.55		
$Score(c_1)$	6.95	$Score(c_2)$	1.74

図 3.4: Yarowsky のモデルによる WSD の例

3.2 Yarowsky の手法の拡張

3.2.1 コロケーションの導入

Yarowsky のモデルでは，WSD に用いる特徴として，周辺に出現する自立語のみを用いていた．以下，この特徴を BOW 特徴 (Bag-of-Words 特徴) と呼ぶ．BOW 特徴は，文脈

中での単語の出現のみを考慮し，単語の並び方などを考慮しない特徴の抽出方法である．しかし，BOW 特徴以外にも，対象語の直前・直後に出現する単語や，対象語と統語的関係 (主語-動詞，目的語-動詞，など) にある単語が WSD の有効な手がかりになることが知られている [7]．特に，動詞を対象とした WSD については，BOW 特徴よりも対象語の直前・直後に出現する単語が語義を決めるための有力な手がかりになると考えられる．

本論文では，対象語 (語義を決めたい単語) 及びその直前・直後に出現する単語の列をコロケーション特徴と呼び，BOW 特徴に加えて，これを Yarowsky のモデルにおける特徴として利用する手法を提案する．コロケーション特徴の定義を図 3.5 に示す．t は対象語，w_i は対象語から見て相対位置 i に出現する単語を表す．1 つの対象語から 4 種類のコロケーション特徴を抽出する．

$$
\begin{aligned}
w_{-2}{+}w_{-1}{+}(t)\\
w_{-1}{+}(t)\\
(t){+}w_1\\
(t){+}w_1{+}w_2
\end{aligned}
$$

図 3.5: コロケーション特徴

図 3.6 は,「とろ火で味を (吸う) 工程が大切だ。」という例文から抽出されるコロケーション特徴の具体例である．この例では (吸う) が対象語である．

味+を+(吸う)
を+(吸う)
(吸う)+工程
(吸う)+工程+が

図 3.6: コロケーション特徴の例

BOW 特徴では対象語を区別しないで特徴を抽出しているのに対し，コロケーション特徴では対象語を区別して抽出している，つまり t もコロケーション特徴に含めていることに注意していただきたい．これは，コロケーションは対象語に強く依存するという観察に基づく．例えば，図 3.1 の例では，分類項目 2.1533 のサブコーパスにおいて，対象語が「吸う」の例文の文脈と「注ぎ足す」の例文の文脈に「とろ火」が出現するため，BOW 特徴として「とろ火」が抽出され，またそのスコアが算出される．直観的には，BOW 特徴「とろ火」のスコアは,「とろ火」が分類項目 2.1533 とどの程度関連性が強いか (2.1533 の意味を持つ単語の周辺にどれだけ現れやすいか) を表す．同様に，図 3.1 の例では，対象語を区別しないでコロケーション特徴を抽出すると，対象語が「吸う」の例文からは「味+を」というコロケーション特徴が抽出される．しかし,「味+を」というコロケーションが分類項目 2.1533 の意味を持つ全ての単語の直前に現れやすいとは限らない．例えば,「にじむ」も分類項目 2.1533 の意味を持つが,「味+を+(にじむ)」は不自然な単語の並びであ

る．したがって，本研究では，コロケーション特徴に対象語自身を含める．すなわち，対象語が「吸う」の例文からは「味+を+(吸う)」を，対象語が「注ぎ足す」の例文からは「煮汁+を+(注ぎ足す)」を，対象語が「にじむ」の例文からは「で+怒られて+(にじむ)」をコロケーション特徴として抽出する．

コロケーション特徴は，厳密には分類項目の特徴を表すものではなく，分類項目における特定の単語の特徴を表すものである．上記の例で言えば,「味+を+(吸う)」は，分類項目 2.1533 における「吸う」の特徴を表す．

本研究の拡張モデルでは，BOW 特徴もコロケーション特徴も同じように取り扱う．サブコーパスを作成後，BOW 特徴とともにコロケーション特徴を抽出する．また，コロケーション特徴の重みは BOW 特徴と同じく式 (3.4) のように計算する．

3.2.2 単義の単語のみの利用

3.1.2 項において図 3.2 の例を用いて説明したように，Yarowsky のモデルでは，単語の多義性を考慮しないため，分類項目の特徴として誤ったものが獲得されたり，特徴のスコアの信頼性が低いという問題がある．これに対し，本研究では，分類項目毎にサブコーパスを作成する際に，単義の単語，すなわち一つの分類項目にしか登録されていない単語のみを利用する手法を提案する．すなわち，分類語彙表から多義語をあらかじめ除去し，その後 Yarowsky のモデルを学習する．サブコーパスの量は減少するが，誤った特徴が抽出されなくなること，特徴のスコアの信頼性が向上することが期待できる．

単義の単語のみを訓練データとして用いるとき，コロケーション特徴は利用できない．本研究におけるコロケーション特徴は WSD の対象語 t を含む．単義の単語のみを訓練データとしたとき，対象語 t は常に単義である．したがって，学習されたモデルにおいて，コロケーション特徴の t は常に単義の単語である．一方，学習したモデルを WSD に実際に適用する際には，対象語は多義語である．単義の単語は語義 (分類項目) を一つしか持たないため，WSD モデルを用いて語義を決定する必要がないためである．したがって，コロケーション特徴自体は単義の単語のみを訓練データとしたときでも獲得できるが，それはテスト文の文脈に出現することはなく，語義の曖昧性を解消する際，式 (3.4) のスコアの計算に使われることはない．そのため，単義の単語のみを訓練データとするときは BOW 特徴のみを用いる．

3.2.3 訓練データの漸進的増加

3.2.2 項で述べた手法の問題点は，単義の単語のみを用いることで訓練データに用いる用例の数が減少することである．この問題を解決するために，bootstrapping の手法を適用し，訓練データを漸進的に増加させることを試みる．また，訓練データを増加させるための手法として,「語義推定モデル」と「語義絞り込みモデル」の 2 つを提案する．以下，その手続きを述べる．

1. 初期の WSD モデル M_1 を学習する．単義の単語のみを訓練データとして用いる．

2. 訓練データにおける多義語に対し，モデル M_j を適用する．多義語が属する分類項目 c_i に対して式 (3.4) のスコア $score(c_i)$ を求める．

3. WSD モデルを再学習する．学習データとして，単義語と多義語の両方を用いる．

   (3-1) **語義推定モデル**
   多義語の分類項目として $score(c_i)$ が最大となるものをひとつ選択する．
   そのスコアが閾値 T_1 以上の多義語のみを学習データとする．候補となっている分類項目からひとつ選択することで，単義の単語とみなす.

   (3-2) **語義絞り込みモデル**
   多義語の分類項目のうち $score(c_i)$ が閾値 T_2 以上のものについて，分類項目の特徴を獲得する．多義語の意味はひとつには決めないが，信頼性の低い語義の特徴を学習しないようにする．

   得られたモデルを M_{j+1} とする．

4. Step 2.〜3. を繰り返す.

語義推定モデルのフローチャートを図 3.7 に示す．この手法では，WSD モデル M_j を訓練データに適用後，判定の信頼度が十分高い用例についてはその対象語の語義を一つ選択し，以後それを単義の単語として訓練データに追加し，新しい WSD モデル M_{j+1} を学習する．語義推定モデルにおける多義語の処理を下記の例を用いて説明する．

(例) … 財政 の 悪化 に （繋がる） 恐れ が 有る …

「繋がる」は 2.1110，2.1131，2.1560 の 3 つの分類項目に属する多義語である．いま，語義の分類モデル M_1 を適応し,「繋がる」の各分類項目のスコアが表 3.2 に示す値になったとする．閾値 T_1 が 10 のとき，1 位のスコアは T_1 よりも大きい．したがって，この用例における「繋がる」は，高い確信度で語義を 1 位の分類項目 2.1110 に決めることができるとみなす．新しいモデル M_2 を学習する際には，この用例における「繋がる」を分類項目 2.1110 の語義を一つだけ持つ単義の単語として扱い，この用例の文脈から 2.1110 の特徴を獲得する．

表 3.2: 「繋がる」の分類項目のスコアの計算例

分類番号	分類項目	スコア
2.1110	関係	17.55
2.1131	連絡・所属	12.23
2.1560	接近・接触・隔離	8.07

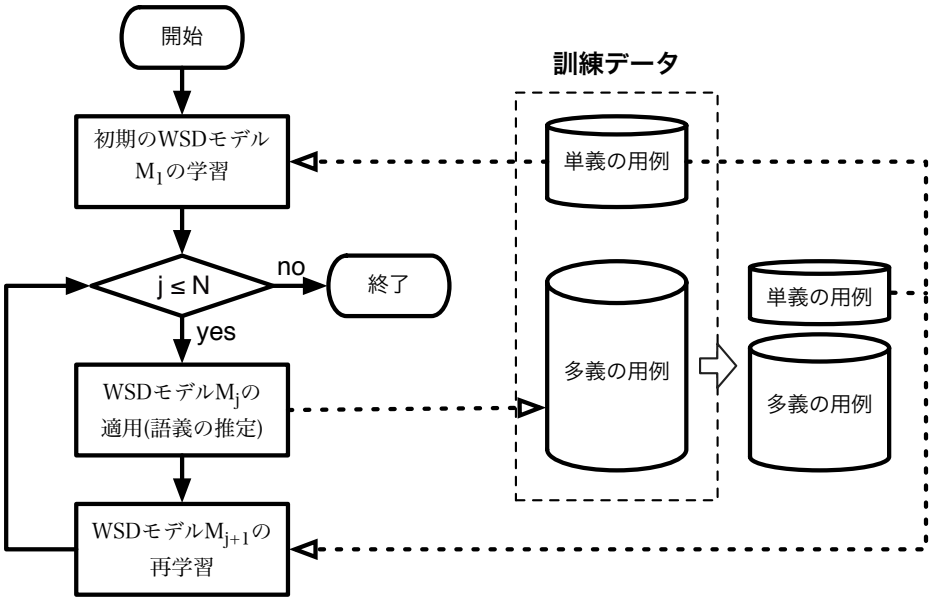

図 3.7: 語義推定モデル

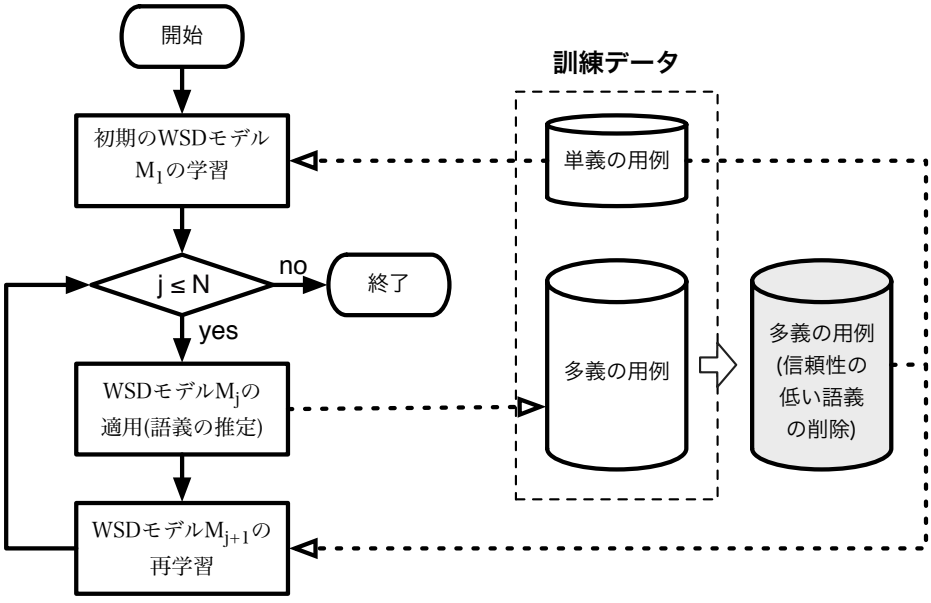

図 3.8: 語義絞り込みモデル

語義絞り込みモデルのフローチャートを図 3.8 に示す．この手法では，WSD モデル M_j を訓練データに適用し，スコア $score(c)$ が閾値 T_2 より小さい分類項目は正しくない語義である可能性が高いとみなす．新しい WSD モデル M_{j+1} を学習する際には，正しくない分類項目については特徴を獲得しない．例えば，前述の「財政の悪化に (繋がる) 恐れがある」という用例に対し，WSD モデル M_1 を適用し,「繋がる」の 3 つの分類項目におけるスコアが表 3.2 のように計算されたとする．閾値 T_2 が 10 のとき，これより小さいスコアを持つ分類項目 2.1560 は，この用例における「繋がる」の語義に該当しないとみなす．新しい WSD モデル M_2 を学習する際には，この用例の文脈から，分類項目 2.1110 と 2.1131 の特徴は獲得するが，不正解とみなした分類項目 2.1560 の特徴は獲得しない．

語義絞り込みモデルは，厳密には訓練データの量を漸進的に増加させる手法ではなく，不正解の語義を除外することにより，訓練データの品質を徐々に向上させていく手法である．しかし，語義推定モデルも語義絞り込みモデルも，図 3.7 と図 3.8 に示すように，ほぼ同じ処理の流れで WSD モデルを学習する．そのため，本論文では，語義絞り込みモデルも訓練データを漸進的に増加させる手法のひとつとして扱う．

語義推定モデルにおける閾値 T_1 は，訓練データに追加する例文の語義の正確性と，訓練データに追加する例文の量をコントロールする働きをする．一般に，追加する例文の語義の正確性と例文の量はトレードオフの関係にある．T_1 を高く設定すれば，WSD の判定結果の信頼性が増すが，語義を決定する事例の数が減るため，訓練データの増分は小さくなる．逆に T_1 を低く設定すれば，多くの多義語に対して語義を決定することになり，訓練データを急速に増やすことができるが，自動推定した語義が誤りである可能性も高くなる．語義絞り込みモデルにおける閾値 T_2 も同様の働きをする．T_2 を高く設定すれば，正しくない語義のみを除外できる可能性が高くなるが，多義語から除外できる語義の数は少なくなる．逆に T_2 を低く設定すれば，多義語からより多くの語義を除外することができるが，正しい語義を誤って除外する可能性も高くなる．したがって，閾値 T_1 と T_2 は，自動推定する語義の正確性と訓練データ量の変化の大きさのバランスを考慮して決定する必要がある．本研究では，閾値 T_1 と T_2 は実験的に決定する．詳細は 4.2 節で述べる．

3.3 分類項目の辞書引き

本節では，提案手法の実装，特に分類語彙表の辞書引きの詳細について述べる．分類語彙表の辞書引きとは，入力として与えられた単語に対し，その単語が登録されている分類項目を全て取得する処理を指す．提案手法では，分類項目の辞書引きは何度も行われ，重要な役割を果たす．3.1.1 項のサブコーパスの作成では，コーパスに出現する単語に対し，分類項目を辞書引きし，その結果得られた分類項目のサブコーパスにその単語と前後 20 単語からなる用例を追加する．学習した語義の分類モデルをテスト文に適用する際には，対象語に対して分類項目の辞書引きを行い，候補となる分類項目 (語義) のリストを得る．

分類語彙表では，動詞や形容詞などの活用語は全て基本形で登録されている．したがって，辞書引きの際には基本形で分類語彙表を検索する必要がある．また，分類語彙表では

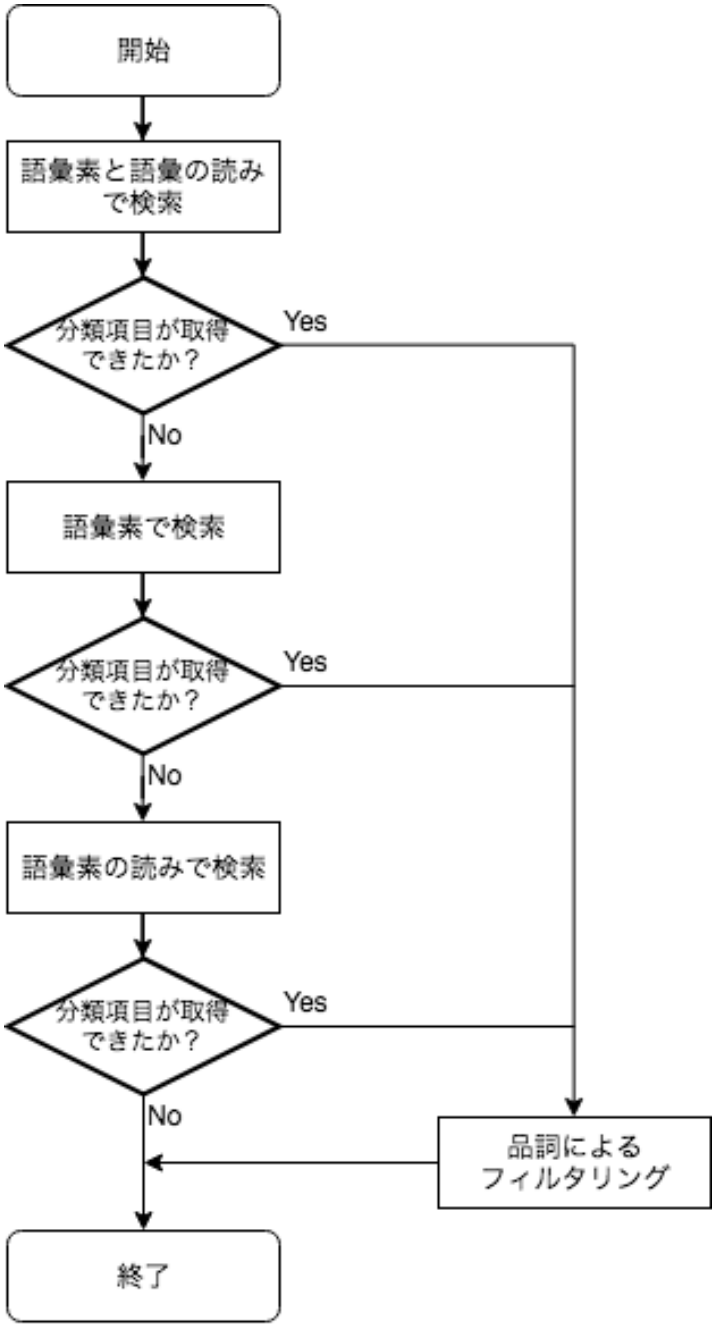

図 3.9: 分類項目の辞書引きの手続き

単語の基本形の読みも登録されている．以下，BCCWJ での用語にあわせて，単語の基本形を「語彙素」，基本形の読みを「語彙素の読み」と記す．

図 3.9 に辞書引きアルゴリズムのフローチャートを示す．分類項目を辞書引きしたい単語が入力されたとき，まずその語彙素と語彙素の読みを得る．分類モデルの学習には BCCWJ を用いたが，BCCWJ では全ての単語に対して語彙素と語彙素の読みが与えられている．一方，語義の分類モデルを未知の文に適用する際にも対象語の語彙素と語彙素の読みが必要である．4 節で述べる評価実験では，テスト文も BCCWJ(ただし，訓練データとは異なる文) を用いるため，語彙素と語彙素の読みも与えられている．第一段階では語彙素と語彙素の読みで検索をする．この段階で一つも分類項目を得ることができない場合，第二段階に移行する．第二段階では，語彙素のみで検索をする．第二段階で分類項目を得ることができない場合，第三段階に移行する．第三段階では，語彙素の読みのみで検索を行う．この段階で分類項目を 1 つも得ることができなかった場合は辞書引きに失敗することになる．

分類語彙表から 1 つ以上の分類項目が検索された場合，品詞によるフィルタリングを行う．分類項目には対応する 5 桁の分類番号が付与されている．また，分類番号の先頭の数字は 1 から 4 に分かれている．表 3.3 に示すとおり，分類語彙表では「類」を最も基本的な分類とし,「体の類」「用の類」「相の類」「その他の類」の 4 つの類がある.「体の類」は名詞を,「用の類」は動詞を,「相の類」は形容詞を,「その他の類」はそれ以外の品詞を表わす．また，類は分類項目の分類番号の最初の一桁の数字で区別されている．1 は「体の類」，2 は「用の類」，3 は「相の類」，4 は「その他の類」を表わす．なお，表 3.3 では，参考のため分類語彙表の「部門」の一覧も載せている．部門とは類の下の分類であり，分類項目の分類番号の最初の 2 桁の数字で表わされる．例えば,「1.1 抽象的関係」は「1 体の類」の下位に属する部門である．

表 3.3: 類と部門の関係

1 体の類	2 用の類	3 相の類
1.1 抽象的関係	2.1 抽象的関係	3.1 抽象的関係
1.2 人間活動の主体		
1.3 人間活動－精神および行為	2.3 精神および行為	3.3 精神および行為
1.4 生産物および用具		
1.5 自然物および自然現象	2.5 自然現象	3.5 自然現象

品詞フィルタリングでは，分類項目の「類」の分類に基づき，辞書引きする単語の品詞と分類項目の品詞が一致しているかをチェックし，一致していない分類項目を除外する．なお，BCCWJ では，全ての単語に対してその品詞の情報が与えられている．まず，辞書引きする単語の品詞が「名詞」または「代名詞」のときは,「体の類」に相当する分類項目のみ，つまり分類番号が 1 で始まる分類項目のみを残す．品詞が「動詞」のとき,「用の類」に相当する分類項目のみ，つまり分類番号が 2 で始まる分類項目のみを残す．品詞が接尾辞のとき，分類語彙表では「体の類」と「相の類」に接尾辞に相当する単語が登録さ

れていることがあったので，分類番号が 1 または 3 で始まる分類項目のみを残す．品詞がそれ以外のとき，名詞または動詞の分類項目は不適切なので，分類番号が 1 または 2 で始まる分類項目を除外する．なお，形容詞のほとんどは「相の類」に属する分類項目にしか登録されていなかったので，フィルタリング条件は設定しなかった．以上のフィルタリング条件を表 3.4 にまとめる．

表 3.4: 品詞フィルタリングの条件

品詞	処理内容
名詞または代名詞	分類番号が 1 で始まる分類項目のみを残す
動詞	分類番号が 2 で始まる分類項目のみを残す
接尾辞	分類番号が 1 または 3 で始まる分類項目のみを残す
上記以外	分類番号が 1 または 2 で始まる分類項目を除外する

第4章　評価実験

本章では，3 章で説明した提案手法の評価実験について述べる．4.1 節では，実験データや評価基準などの実験設定について述べる．4.2 節では，3.2.3 項で述べた手法におけるパラメータを決めるための予備実験について述べる．最後に，4.3 節では実験結果を報告し，その考察を述べる．

4.1　実験設定

4.1.1　実験データ

実験データとして，BCCWJ に対して分類語彙表の分類項目を人手で付与したコーパス [2] を用いる．実験データにおける WSD の対象単語数および一単語当たりの語義数の平均を表 4.1 に示す．対象語ののべ数は 3,912 語，異なり数は 924 語である．実験データにおける対象語の品詞は,「名詞」「動詞」「その他」に大きく分けることができる.「その他」は感動詞，形状詞，形容詞，接尾辞，代名詞，副詞，連体詞のいずれかの品詞に該当する．表 4.1 に示すように，実験データでは名詞の割合がおよそ 63%と一番多い．次に多いのは動詞で，その割合は 30%である．以降で実験結果を報告するときは，実験データ全体の結果の他に,「名詞」「動詞」「その他」に対する結果も報告する．

対象語は全て語義を 2 つ以上もつ多義の単語である．一単語あたりの語義数が 3.12 であることから，ランダムに語義を選択したときの正解率はおよそ 32%となる．

表 4.1: 実験データ

品詞	名詞	動詞	その他	合計
対象語数	2467	1175	270	3912
平均語義数	2.40	4.82	2.28	3.12

4.1.2　評価基準

WSD モデルは正解率と適用率で評価する．正解率は，WSD モデルによって語義を決定した対象語のうち，正しく語義を推定することができた対象語の割合である．その定義

を式 (4.1) に示す.

$$\text{正解率} = \frac{\text{正解の語義を選択できた対象語の数}}{WSD\text{手法によって語義を決定した対象語の数}} \tag{4.1}$$

適用率は，語義を決定できる対象語の割合である．その定義を式 (4.2) に示す.

$$\text{適用率} = \frac{\text{語義を決定できた対象語数}}{\text{テストデータにおける対象語の総数}} \tag{4.2}$$

Yarowsky のモデルも，本研究の提案手法も，対象語が持つ分類項目 c_i に対し，そのスコア $score(c_i)$ を計算し，それが最大の c_i を選択する．このとき，1 位のスコアの値が十分に高くないときは，判定の信頼性が低いと判断し，語義を決定しないこととする．すなわち，1 位のスコアがあらかじめ決められた閾値 T 以上なら語義を決定し，T より小さければ語義を決定しない．適用率は，WSD モデルがどれだけ多くの対象語の語義を決めることができるかを評価する指標である.

今回の実験では正解率を主な指標として WSD モデルを評価する．適用率は 4.2 項の予備実験で提案手法のパラメータを決定する際に使用する.

4.2 予備実験

3.2.3 項で述べた訓練データを漸進的に増加させる手法のパラメータを決定する予備実験を行った．具体的には，語義推定モデルにおける閾値 T_1 と，語義絞り込みモデルにおける閾値 T_2 を決めるための実験を行った．これらの閾値は，訓練データを漸進的に増加する過程において，自動推定する語義の正確性と訓練データ量の変化の大きさのバランスを考慮して決定する必要がある.

Yarowsky のオリジナルの手法について，語義を決定するか否かを決めるスコアの閾値 T を変化させ，正解率と適用率の変化を調べた．結果を図 4.1〜4.4 に示す．これらの図において，横軸は閾値 T，縦軸は正解率ならびに適用率を表す．図 4.1 はテストデータ全体 (全品詞) の結果，図 4.2，図 4.3，図 4.4 はそれぞれ名詞，動詞，その他の品詞の結果を示している.

正解率と適用率はトレードオフの関係にある．閾値 T を大きく設定すれば，1 位の分類項目のスコア $score(c_i)$ が十分に大きいときのみ語義を決めるため，正解率は高くなる．一方，1 位のスコアが閾値 T を越えて語義を決めることができるテスト文の数は減るため，適用率は低くなる．全品詞 (図 4.1) と名詞 (図 4.2) については，閾値が 18 までは正解率がほぼ一定であるが，その後急激に上昇し，名詞の場合は閾値が 22 になると正解率が 1 に到達する．これに対し，動詞 (図 4.3) やその他の品詞 (図 4.4) については，閾値を十分に大きくしても正解率は大幅に上昇するわけではない．一方，適用率はどの図でも単調に減少している.

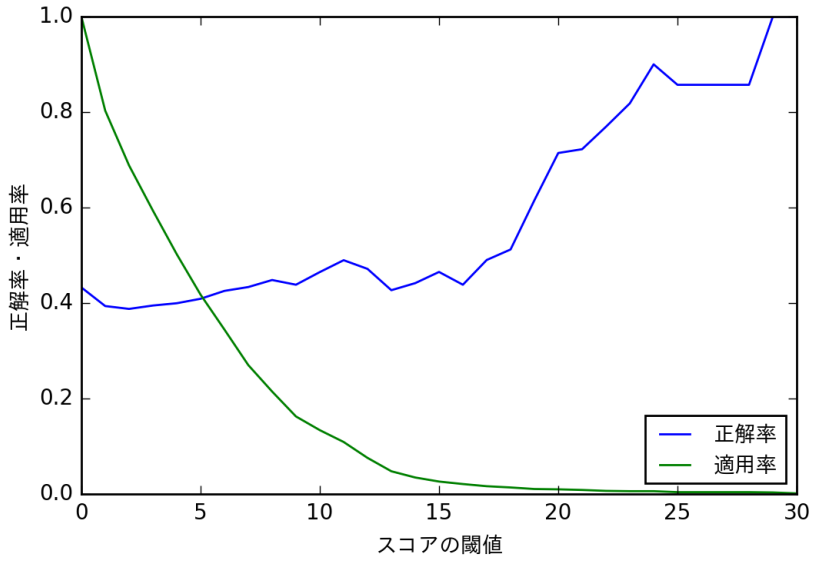

図 4.1: Yarowsky のモデルの正解率と適用率 (全品詞)

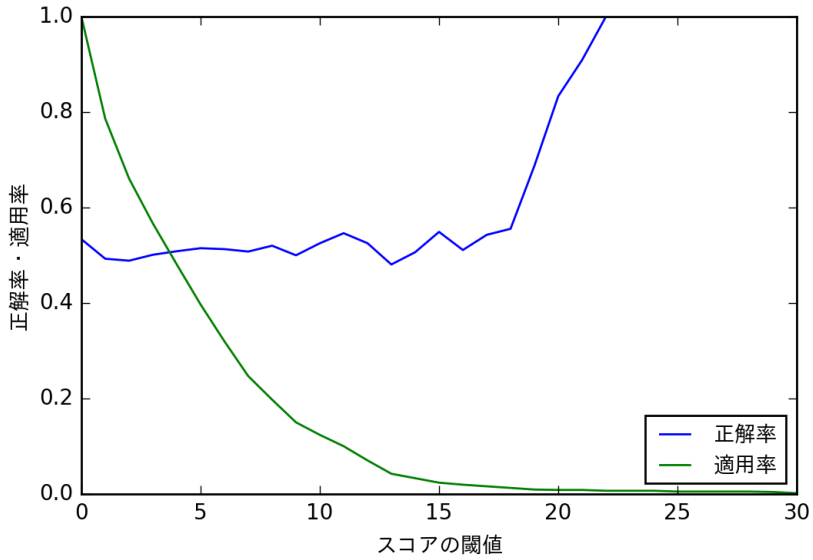

図 4.2: Yarowsky のモデルの正解率と適用率 (名詞)

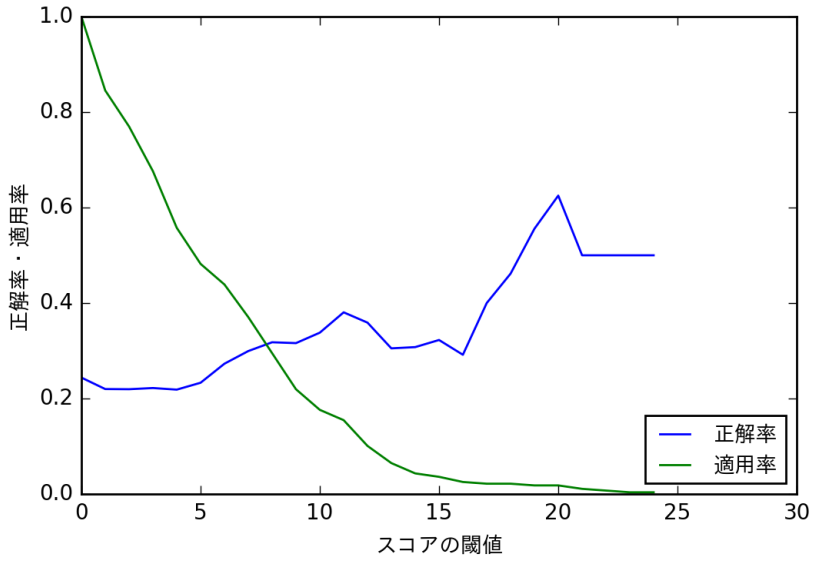

図 4.3: Yarowsky のモデルの正解率と適用率 (動詞)

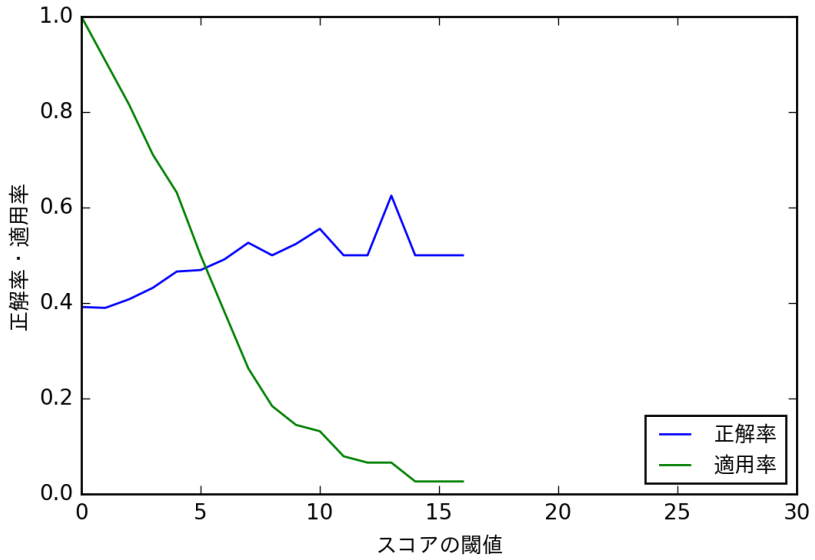

図 4.4: Yarowsky のモデルの正解率と適用率 (その他)

訓練データを漸進的に増加させる手法における閾値 T_1 と T_2 も，図 4.1〜図 4.4 の横軸の閾値 T と同様の働きをする．今回の実験では，自動推定した語義の信頼性がある程度高くなるように，WSD の正解率がおよそ 80%になる閾値を選んだ．具体的には，図 4.1 において，$T = 23$ のときに正解率が 81.8%となり，ここで初めて 80%を越えるため，$T_1 = T_2 = 23$ と設定した．このときの適用率は 0.59%である．これは訓練データにおける多義語のうち 0.59%しか語義を推定することができず，1 回の反復で増やすことのできる訓練データの量はわずかであることを意味する．しかし，誤った分類項目の特徴が学習されることの弊害の方が大きいと考え，WSD の正解率が十分高くなるように閾値を設定した．

上記の閾値の決め方は以下の問題がある．まず，語義推定モデルの閾値 T_1 と語義絞り込みモデルの閾値 T_2 は，本来は最適値が異なると考えられるが，今回の実験では同じ値に設定した．また，図 4.2，図 4.3，図 4.4 に示されている通り，正解率と適用率の変化は品詞によって異なる．名詞の場合は $T = 15$ のときに正解率が 0.8 に達する．一方，動詞やその他の品詞のときは，閾値を十分高く設定しても正解率は 0.8 に達しないため，別の基準で閾値を選択する必要がある．このように，名詞，動詞，その他の品詞のそれぞれについて，閾値 T_1 や閾値 T_2 を決めるべきであろう．しかし，今回の実験では品詞によって閾値を変えることはせず，全ての品詞について同じ閾値を用いた．さらに，閾値 T_1 と T_2 の決定に用いたデータと，次節で報告する WSD モデルの評価には，ともに表 4.1 に示した実験データを用いている．しかし，T_1 と T_2 のようなパラメータは，本来はテストデータとは別の開発データを用いて最適化するべきである．最後に，図 4.1〜4.4 に示した Yarowsky のモデルは，研究の途中段階で実装したものであり，次節で報告する Yarowsky のモデルとは実験結果が異なる．また，本研究で提案する訓練データの漸進的増加手法における初期モデル M_1 は，単義の単語のみを訓練データとして用いるモデルである．したがって，Yarwosky のオリジナルのモデルではなく，単義の単語のみを使う初期モデルの正解率と適用率の変動を調べ，閾値 T_1 と T_2 を決定するべきである．これらの問題の解決は今後の課題である．

4.3 実験結果

まず，以下の WSD モデルを評価する．

- Yarowsky のオリジナルのモデル
- コロケーション特徴を用いたモデル (3.2.1 項)
- 単義の単語のみを訓練データとしたモデル (3.2.2 項)

なお，3.2.2 項で説明したように，単義の単語を訓練データとしたときはコロケーション特徴を用いることはできないため，BOW 特徴のみを用いる．各モデルの学習には BCCWJ の全ての文を用いた．

実験結果を表 4.2 に示す．M は単義の単語を，P は多義の単語を訓練データとして用いたことを表す．また，BOW と COL はそれぞれ BOW 特徴とコロケーション特徴を用いたことを表す．Yarowsky のオリジナルのモデルは表 4.2 の 2 行目に相当する．

従来の BOW 特徴に加え，コロケーション特徴を追加することの効果について考察する．動詞について,「M+P BOW+COL」と「M+P BOW」を比較すると，前者は後者と比べて正解率が 3.4 ポイント高い．したがって，コロケーション特徴は動詞の WSD に有効に働くと言える．ただし，名詞について比較すると，正解率が 1.0 ポイント低下した．また，その他の品詞について 2 つのモデルを比較すると，コロケーション特徴を導入することで正解率が 0.3 ポイント低下した．テストデータ全体でも，コロケーション特徴を使わないモデルの方が正解率が 0.2 ポイント高く，コロケーション特徴を使用することで正解率はわずかに低下している．以上をまとめると，動詞については正解率が向上するが，それ以外については低下する．このため，全体の正解率を向上させるためには，品詞ごとに用いる特徴を変更する方法が考えられる．例えば，BOW 特徴だけを用いた WSD モデルと，BOW 特徴とコロケーション特徴の両方を用いた WSD モデルを学習し，対象語が動詞のときには後者のモデルを，それ以外の品詞のときは前者のモデルを適用することで，全体の正解率を向上させることが期待できる．

一方，単義の単語のみを訓練データとして用いた WSD モデル「M BOW」と「M+P BOW」の正解率を比較すると，名詞では 4.3 ポイント，動詞では 7.5 ポイント，その他の品詞では 3.0 ポイント，テストデータ全体では 4.8 ポイントの向上が確認できた．単義のみの単語を訓練データとして用いることで，全ての品詞について正解率が大きく向上した．

表 4.2: 実験結果 (全コーパス)

データ	特徴	名詞	動詞	その他	全て
*M+P	BOW	0.555	0.409	0.433	0.519
M+P	BOW+COL	0.545	0.443	0.430	0.517
M	BOW	0.598	0.484	0.463	0.567

* (Yarowsky 1992) に相当

次に，3.2.3 項で述べた訓練データの漸進的増加法を評価する．4.2 節の予備実験により，閾値 T_1 と T_2 は 23 と設定した．また，反復回数は 1 回のみとした．結果を表 4.3 に示す．B1 は「語義推定モデル」を，B2 は「語義絞り込みモデル」を表す．ただし，今回の実験では，実装上の問題から，3.2.3 項で述べた訓練データの漸進的増加手法の Step 2 において，BCCWJ 全体のうち 44%のテキストについてしか初期モデルによって語義を推定することができなかったため，本論文ではこの部分コーパスのみを訓練データとして用いた結果を報告する．表 4.4 は，BCCWJ を構成するファイルのうち，今回の実験に用いた部分コーパスに含めたファイルおよび含めなかったファイルを示している．また，表 4.5 に部分コーパスに含まれる単語数およびコーパス全体に対する割合を示す．参考のため，表 4.2 に示した 3 つの WSD モデルについても，部分コーパスを用いて学習し，これをテス

トデータに適用したときの正解率を示す.

表 4.3: 実験結果 (部分コーパス)

データ	素性	名詞	動詞	その他	全て
M+P	BOW	0.544	0.264	0.456	0.451
M+P	BOW+COL	0.501	0.277	0.421	0.426
M	BOW	0.535	0.432	0.461	0.499
M+B1	BOW	0.545	0.443	0.467	0.517
M+B2	BOW+COL	0.536	0.462	0.451	0.509

語義推定モデルである「M+B1 BOW」と「M BOW」を比較すると，初期モデルで語義を推定した用例を訓練データに追加することで，テストデータ全体の正解率は 49.9%から 51.7%に向上した．品詞別にみると，名詞は 1 ポイント，動詞は 1.1 ポイント，その他の品詞は 0.6 ポイント向上した．いずれの品詞も正解率は向上したが，対象語の品詞と正解率の差の相関については特に顕著な傾向は見られなかった．一方，語義絞り込みモデルである「M+B2 BOW」と「M BOW」を比較すると，初期モデルによって信頼性が低いと判定した語義を除外することで，テストデータ全体の正解率は 49.9%から 50.9%に向上した．品詞別にみると，名詞は 0.1 ポイント，動詞は 3 ポイント向上したが，その他の品詞は 1 ポイント低下した．語義絞り込みモデルは特に動詞に有効に働くことがわかった．以上の結果から，訓練データを漸進的に増加させる提案手法の有効性が確認された．また，語義推定モデルと語義絞り込みモデルを比較すると，動詞では語義絞り込みモデルの方が正解率が高いが，それ以外の品詞では語義推定モデルの方が正解率が高く，テストデータ全体では語義推定モデルの正解率は語義絞り込みモデルよりも 0.8 ポイント高かった．

「M+P BOW」「M+P BOW+COL」「M BOW」の 3 つの WSD モデルについては，全コーパスを用いた表 4.2 の実験結果とおおむね同じような傾向が見られた．ただし，名詞について「M+P BOW」と「M BOW」を比較すると，表 4.2 とは異なり，前者のほうが正解率が高い．この原因として，表 4.3 の実験では全体の 46%の量しか訓練データに用いておらず，また単義の単語のみを用いる場合は特徴抽出に利用できる例文数がさらに減るため，訓練データ量の減少が特に大きく影響していると考えられる．また，Yarowsky のモデル「M+P BOW」の動詞の正解率は，表 4.2 では 40.9%だったのに対し，表 4.3 では 26.4%と大きく低下している．同様に，提案手法「M+P BOW+COL」の動詞の正解率は，表 4.2 では 44.3%，表 4.3 では 27.7%である．単義と多義の単語の両方を訓練データとして用いる WSD モデルにおいて，訓練データの量は動詞の WSD の正解率に大きく影響することがわかった．

表 4.4: 訓練データの漸進的増加手法の評価に用いた BCCWJ のファイル

文章種類	ファイル名	単語数	使用の有無
書籍 (出版)	PB1	6909546	無
	PB2	6980662	無
	PB3	6883203	無
	PB4	7021168	無
	PB5	5897002	無
雑誌	PM	5390926	有
新聞	PN	1615076	有
書籍 (図書館)	LBe	1473538	有
	LBr	2065348	有
	LBo	2210679	有
	LBa	834579	有
	LBg	1624001	有
	LBd	1278036	有
	LBc	1084515	有
	LBl	2272786	有
	LBi	1927365	有
	LBf	1537172	有
	LBm	2090728	有
	LBp	1921401	有
	LBs	2165114	有
	LBb	1034586	有
	LBk	2018516	有
	LBj	2043037	有
	LBh	1805712	有
	LBn	2154028	有
	LBt	2108468	有
	LBq	2154481	有
白書	OW	5693403	無
教科書	OT	1125388	有
広報紙	OP	4697015	有
ベストセラー	OB	4434404	有
Yahoo!知恵袋	OC	12066093	無
Yahoo!ブログ	OY	13067279	無
韻文	OV	233457	有
法律	OL	1206120	有
国会会議録	OM	5599178	無

表 4.5: 訓練データの漸進的増加手法の評価に用いた単語数

使用の有無	単語数	割合
有	54,506,476	44%
無	70,117,534	56%
合計	124,624,010	100%

第5章　おわりに

5.1　本研究のまとめ

本論文では，分類語彙表の分類項目を識別する WSD タスクに対し，Yarowsky のモデルを拡張する手法を提案した．まず，Yarowsky のモデルを分類語彙表の分類項目を語義としたモデルとして実装し，これをベースラインシステムとした．次に，ベースラインモデルを拡張する 3 つの手法を提案した．

一つ目は，コロケーション特徴の導入である．対象語ならびにその直前・直後に出現する単語列を新たな特徴として用いた．Yarowsky の手法では，特徴に BOW を用いており，個々の単語の情報は WSD に用いられるが，単語間の関連性を WSD に用いることはできない．しかし，対象語の直前・直後にある単語もまた特に動詞の WSD に有効とされていることから，コロケーション特徴を導入した．実験の結果，動詞についてはコロケーション特徴を利用することで正解率が 3.4 ポイント向上したことを確認した．

二つ目は，単義の単語のみを学習データに利用する方法である．Yarowsky の手法では，コーパスから分類項目の特徴を獲得する際には，多義の単語の文脈からも自立語を抽出して特徴とする．このとき，多義の単語が持つ複数の分類項目のうち，その文脈で使われている正しい分類項目はひとつだけであり，残りの分類項目は誤りであるが，誤りの分類項目については正しくない特徴が獲得されることになり，結果としてコーパスから獲得される分類項目の特徴やその特徴のスコアの信頼性が低くなるという問題がある．単義の単語のみを学習に用いれば，訓練データの量は少なくなるものの，上記の問題を回避できる．評価実験では，単義の単語のみを特徴として利用することで，全ての品詞において正解率が向上し，テストデータ全体に対する正解率が 4.8 ポイント向上したことを確認した．

三つ目は，訓練データの漸進的増加である．まず，単義の単語のみから語義の分類モデルを学習し，それを訓練データの多義語に適用し，判定の信頼度が高い多義語の語義を決定して単義の単語と同じように扱う．この処理を繰り返すことで訓練データの量を少しずつ増やしていく手法を提案した．本研究では，多義語の語義を一つに決める語義推定モデルと，信頼性が低い語義を使用しない語義絞り込みモデルの 2 つを実装した．実験の結果，反復回数を一回としたとき，訓練データを増加させることで正解率が 1.8 ポイント向上した．また，語義推定モデルが語義絞り込みモデルより優れていることがわかった．

上記 3 つの拡張を全て取り入れた語義推定モデルは，ベースラインシステムと比べて，WSD の正解率が 6.6 ポイント向上することを確認した．これにより提案手法の有効性が確認できた．

5.2 今後の課題

本研究の課題について述べる．今回の実験では，訓練データを漸進的に増加させる際，反復回数は 1 としていた．しかし，反復回数を増やすことによって，語義推定モデルではより多くの訓練データが，語義絞り込みモデルでは (多義語が持つ語義の数が少ないという意味で) 良質の訓練データが得られると考えられる．したがって，反復回数を増やしたときに WSD の正解率がどれだけ向上するかを調べる必要がある．また，どれだけ反復回数を増やせば正解率の上昇が飽和するかも調べる．一般に，反復回数を増やすと WSD の正解率は向上するが計算量は増加するというトレードオフがあるが，両者のバランスを考慮して反復回数を設定することも重要な課題である．

今回の評価実験では，コロケーション特徴を用いることにより，動詞の WSD の正解率は向上したが，名詞やその他の品詞の単語については正解率が低下した．したがって，コロケーション特徴は動詞の意味を決める際には有効であるが，動詞以外の品詞の語についてはそれほど有効でないと言える．そこで，動詞については BOW 特徴とコロケーション特徴の両方を用いた分類モデルを，名詞やそれ以外の品詞については BOW 特徴のみを用いた分類モデルを学習する手法が考えられる．対象語の品詞によって，異なる特徴を用いて学習した分類モデルを選択的に適用することで，WSD の正解率を向上させることが期待できる．

また，今回の評価実験では，訓練データを漸進的に増加させる 2 つの手法におけるパラメータを同じ値に設定していた．すなわち，語義識別モデルにおける語義のスコアの閾値 T_1 と，語義絞り込みモデルにおける語義のスコアの閾値 T_2 を同じ値に設定していた．しかし，T_1 と T_2 は同じ値に設定する必要はなく，それぞれ独立に最適化するべきである．また，WSD の判定結果の信頼度を語義のスコアで推定しているが，語義のスコアと信頼度の関係は名詞，動詞，形容詞などの品詞によって異なると考えられる．例えば，語義のスコアが 10 のとき，名詞については WSD の信頼性が十分高いと言えるが，動詞についてはそうではない可能性がある．したがって，T_1 や T_2 といった閾値を品詞別に設定する必要があるだろう．上記も含めて，訓練データの漸進的増加におけるパラメータの最適化は，今後十分に検討する必要がある．

WSD モデルの学習データの増加も今後の課題のひとつとして挙げられる．実験では，実装上の問題により，BCCWJ の全ての文を学習に用いることができなかった．プログラムの効率化によって計算時間を短縮し，BCCWJ の全てのファイルを学習に用いることで，正解率がどれだけ向上するかを調べる必要がある．また，BCCWJ 以外のコーパス，例えばウェブから獲得した大量のテキストを学習データに加えることも検討するべきである．

本研究では，教師なし機械学習の手法を提案した．教師なし機械学習には，大量の学習データを容易に手に入れることができるという利点があるが，分類モデルの精度がそれほど高くないという問題点もある．今回の実験でも，WSD の正解率は最高で 51.7%であり，十分に高いわけではない．これに対し，教師あり機械学習は，正解の分類項目が付与された大量の学習データを作成するためのコストが高いが，WSD の正解率は高い．実際，多くの先行研究で，教師あり機械学習に基づく WSD 手法は優れた成果を残している．両

者の利点を活かすため，これらを併用する WSD モデルを探究することは意義がある．図 5.1 は教師あり機械学習と教師なし機械学習を併用する手法の例を示している．まず，正解の語義 (分類語彙表の分類項目) が付与されていない訓練コーパスから，本研究で提案した Yarowsky の手法の拡張モデルを教師なし学習する．これを訓練コーパスの多義語に適用し，正解の語義を付与して，語義タグ付きコーパスを得る．このコーパスから Support Vector Machine などの教師あり機械学習アルゴリズムによって WSD の分類器を得る．図 5.1 の混合モデルは，Yarowsky の手法の拡張モデルと教師あり機械学習された WSD 分類器を組み合わせて語義の曖昧性を解消する．具体的には，対象単語の分類器が教師あり機械学習できたときには，その分類器を使用し，そうでないときは，Yarowsky の手法の拡張モデルを使用する手法である．このような混合モデルの探究に是非取り組みたいと考えている．

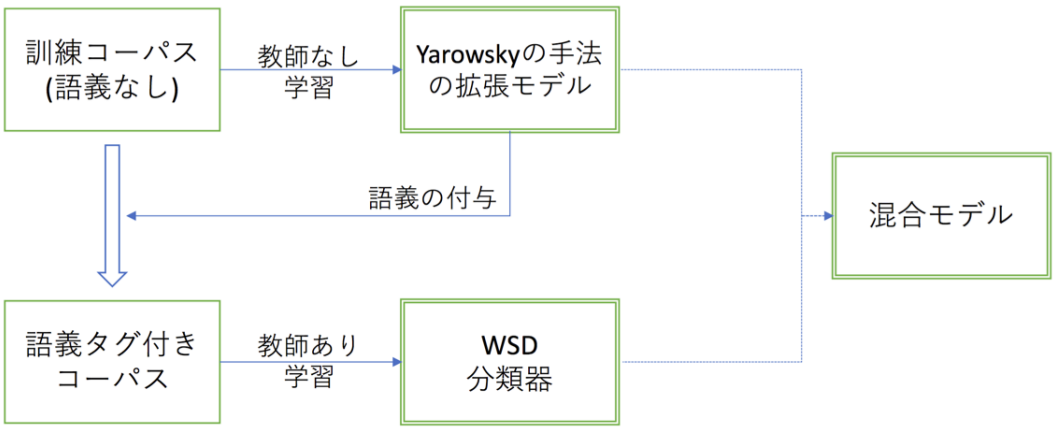

図 5.1: 教師あり機械学習と教師なし機械学習の併用

謝辞

本研究を進めるにあたり，研究の方向性について熱心な御指導を頂きました，白井 清昭准教授に深く感謝するとともに，心より御礼申し上げます．また，本研究に関して多くの有意義なご意見を頂きました東条 敏教授，飯田 弘之教授，池田 心准教授にもこの場を借りて御礼申し上げます．

最後に，これまでの学生生活を支えて頂いた家族に感謝を致します．

参考文献

[1] Yuan Dayu, Richardson Julian, Doherty Ryan, Evans Colin, and Altendorf Eric. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*, 2016.

[2] 加藤祥, 浅原正幸, 山崎誠. 『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション. 言語処理学会第 23 回年次大会発表論文集, pp. 306–309, 2017.

[3] 国立国語研究所. 分類語彙表. 秀英出版, 1964.

[4] 国立国語研究所（編）. 分類語彙表. 大日本図書, 2004.

[5] Kanako Komiya, Minoru Sasaki, Hiroyuki Shinnou, Yoshiyuki Kotani, and Manabu Okumura. Selecting training data for unsupervised domain adaptation in word sense disambiguation. In *Pacific Rim International Conference on Artificial Intelligence*, pp. 220–232. Springer, 2016.

[6] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[7] 村田真樹, 内山将夫, 内元清貴, 馬青, 井佐原均. SENSEVAL2J 辞書タスクでの crl の取り組み. 自然言語処理, Vol. 10, No. 3, pp. 115–133, 2003.

[8] 鈴木類, 古宮嘉那子, 浅原正幸, 佐々木稔, 新納浩幸. 『分類語彙表』の類義語と分散表現を利用した all-words 語義曖昧性解消. 言語処理学会第 23 回年次大会発表論文集, pp. 86–89, 2017.

[9] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING*, pp. 454–460, 1992.