分類語彙表の分類項目を識別する語義曖昧性解消 – Yarowskyモデルの適用と拡張 –

小林 健人　　白井 清昭
北陸先端科学技術大学院大学 先端科学技術研究科
{s1610226,kshirai}@jaist.ac.jp

1 はじめに

語義曖昧性解消 (Word Sense Disambiguation;WSD) は，複数の意味 (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 [3] がある．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．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 (BCCWJ) 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 [1] が，研究者に広く利用できる段階には至っていない．

本論文は，分類語彙表の分類項目を識別することを目的とした教師なし機械学習に基づくWSD手法について述べる [2]．

2 関連研究

分類語彙表は，語を意味によって分類・整理したシソーラスである．分類項目に対し，それに該当する意味を持つ単語が定義されている．例を表1に示す．分類項目には5桁の分類番号が与えられている．分類項目数は895，単語ののべ数は81320である．一つの語が複数の分類項目に分類されることがあり，その語は多義語であるとみなせる．

表 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
親・先祖	1.2120	父, 母, 先祖, ...
哺乳類	1.5501	犬, 猫, うさぎ, ...
鳥類	1.5502	すずめ, はと, かもめ, ...

Yarowskyは，ロジェのシソーラスを語義の定義とした教師なし機械学習に基づく著名なWSD手法を提案した [7]．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．単語を列挙することで語義を定義している点は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．さらに，Yarowskyの手法を問題点を指摘し，これを改良する手法を提案する．

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 [6]．この手法では単語ならびに分類項目の分散表現を基に語義の曖昧性を解消する．評価実験の結果，WSDの正解率は56.3%〜59.6%となった．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．

3 提案手法

3.1 Yarowskyの手法

本項ではYarowskyの手法 [7] を説明する．ただし，分類語彙表を語義の定義に用いることを仮定する．

語義の分類モデルの学習は，図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という2つのステップから構成される．

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コーパスから検索・収集し，サブコーパスを作成する．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BCCWJ[4] を用いた．用例を検索する際には，分類項目に属する単語と基本形が一致する単語を検索し，その前後20単

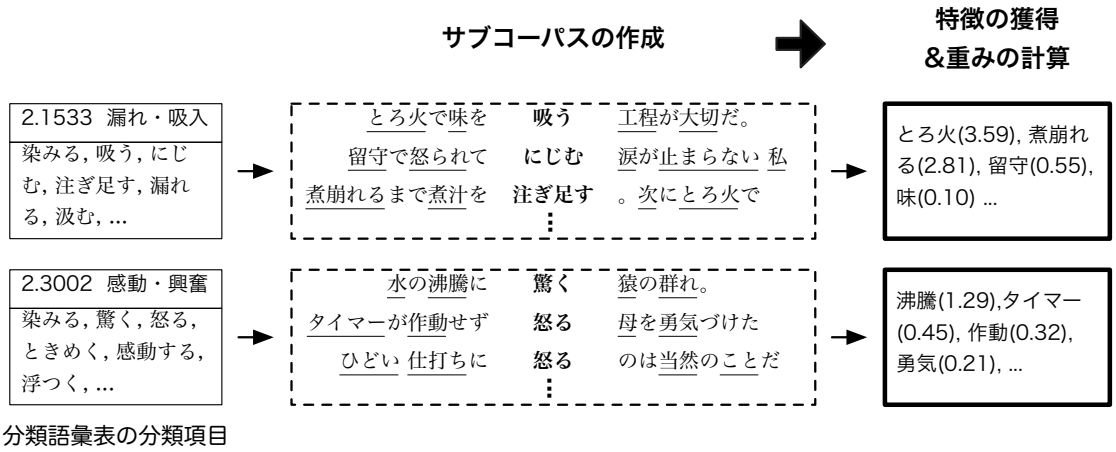

図 1: (Yarowsky 1992) の概要

語とともに用例として抽出する．838 個の分類項目に対し，平均 29,800 件の用例を取得した．

特徴の獲得

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．

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 (1) のように定義される．

$$w(c,f) = \log \frac{Pr(f|c)}{Pr(f)} \tag{1}$$

$Pr(f|c)$ は分類項目 c のサブコーパスにおける f の出現確率，$Pr(f)$ はコーパス全体における f の出現確率である．

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．しかし，Yarowsky は，大量のコーパスから特徴を取得すれば，その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．

本研究では，880 個の分類項目に対し，のべ 8,770,000 の特徴を獲得した．ひとつの分類項目あたりの特徴数の平均は 9,960 である．

語義の推定

分類モデルを学習後，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 (2) 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 (2) における f はテスト文の文脈に出現する特徴を表す．

$$score(c) = \sum_{f\ in\ s} w(c,f) \tag{2}$$

例えば,「染みる」は分類項目として図 1 に示した 2.1533 と 2.3002 を持つ多義語である.「染みる」を含むテスト文が与えられたとき，2.1533 と 2.3002 のそれぞれについて,「染みる」の文脈に出現する特徴に対してその重みの和を求め，その大きい方を選択する．

3.2 Yarowsky の手法の拡張

3.2.1 コロケーションの導入

Yarowskey 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．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 [5]．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，対象語の直前・直後に出現する単語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 2 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．[ ] 内は図 1 における「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．

w_{-2}+w_{-1}+(t)	[味+を+(吸う)]
w_{-1}+(t)	[を+(吸う)]
(t)+w_1	[(吸う)+工程]
(t)+w_1+w_2	[(吸う)+工程+が]

図 2: コロケーション特徴

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．図 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)」は不自然なコロケーションである．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「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．

3.2.2 単義の単語のみの利用

3.1 項で述べたように，Yarowskey 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．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，学習したモデルを実際に適用する際には，対象語は多義語である．したがって，多義語の曖昧性解消にコロケーション特徴を利用することはできない．そのため，単義の単語のみを訓練データとするときは BOW 特徴のみを獲得する．

3.2.3 訓練データの漸進的増加

3.2.2 で述べた手法の問題点は，訓練データに用いる用例の量が減少することである．ここでは，bootstrapping の手法を適用し，訓練データを漸進的に増加させる手法を提案する．この際，語義推定モデルと語義絞り込みモデルの 2 つの手法を考える．

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 (2) 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_{score} 以上の多義語のみを学習データとする．

(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_{score} 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，信頼性の低い語義は学習データから除去する．

得られたモデルを M_{j+1} とする．

4. Step 2.〜3. を繰り返す．

閾値 T_{score} は実験的に決定する．M_1 をテストデータに適用し，$score(c_i)$ が閾値 T 以上のときのみ語義を決定する．T を変動させ，精度の変動を調べる．WSD の精度が 80%になる閾値を T_{score} とする．

4 評価実験

テストデータとして，BCCWJ に対して分類語彙表の分類項目を付与したコーパス [1] を用いる．テストデータにおける 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 2 に示す．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 2: テストデータ

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

各手法の WSD の正解率を表 3 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 3 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したことにより，動詞の正解率が大きく向上した．コロケーション特徴は動詞の WSD に有効であることがわかる．ただし，名詞や他の品詞の正解率は少し下がっている．このため，全体の正解率を向上させるため，動詞のみコロケーション特徴を使うという方式が考えられる．一方，単義の単語のみを訓練データとして用いると，全体的に正解率が大きく向上することが確認できた．

表 3: 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 で述べた訓練データの漸進的増加法を評価する．この実験では T_{score} を 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 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，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[1]．参考のため，表 3 に示した 3 つのモデルを部分コーパスを用いて学習した結果も載せている．

訓練データを漸進的に増加させる方法は，単義の単語のみを用いたモデルに比べて正解率が向上したことが確認された．B1 と B2 を比較すると，B1 の方がわずかに正解率が高かった．

表 4: 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

5 おわりに

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 のモデルを拡張して適用する手法を提案した．本研究で提案する拡張手法は，WSD の正解率を最大で 6.6%(BCCWJ 全体を用いたときは 4.8%) 向上させることが確認できた．

今後の課題として，訓練データにおける漸進的増加において反復回数を増やすこと，名詞および動詞によってモデルや使用する特徴を使い分けること，教師あり機械学習手法と組み合わせることなど，WSD の正解率を更に向上させる方法を探究したい．

参考文献

[1] 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.

[2] 小林健人. 分類語彙表の分類項目を識別する語義曖昧性解消. 修士論文, 北陸先端科学技術大学院大学, 3 2018.

[3] 国立国語研究所（編）. 分類語彙表. 大日本図書, 2004.

[4] Kikuo Maekawa et al. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, 2014.

[5] 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.

[6] 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.

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

[1] 発表時には全コーパスを訓練データとして用いた結果を報告する．