BACTを利用した日本語慣用句意味曖昧性解消

白石 貴大　竹内 孔一
岡山大学大学院自然科学研究科
{shiraishi,koichi}@cl.cs.okayama-u.ac.jp

1　はじめに

近年，インターネットの普及により容易に言語，音声，画像といった情報を発信することができるようになってきている．言語を用いた情報発信では，発信者によって様々な表現が用いられるために同一の意味を持つ文であっても，同一の文となるとは限らない．そこで文を構造化して扱うことができれば自動翻訳や評判分析等，様々な分野に応用することができる．

文の構造化の手法として述語語義と意味役割を用いるもの [1] が提案されている．これらの推定を行う際に慣用句を扱う必要がある．これは慣用句が 2 語以上の単語のまとまりで成ることや，意味曖昧性を持つことによる．例えば図 1 のように，文「彼は床にあぐらをかく」では動詞「かく」を中心に構造化を行う．文「彼は遺産にあぐらをかく」では動詞「かく」ではなく慣用句「あぐらをかく」を中心として構造化を行う必要がある．

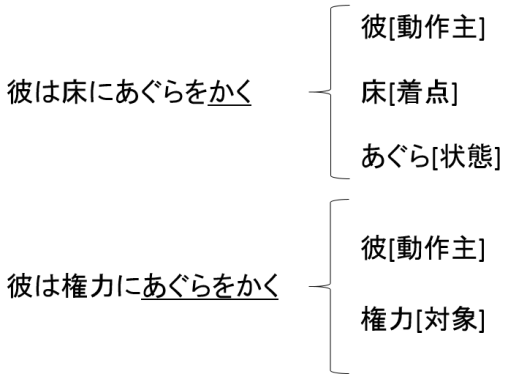

図 1: 構造化の例

先行研究にて SVM を用いた意味曖昧性解消 [2] や異形を吸収した慣用句表現の抽出 [3] が提案されている．橋本ら [2] の手法では SVM を用いた機械学習により正解率 89.19%を得ている．また，守屋ら [3] の手法では，異形に対して大量の事例を準備することは容易ではないという点からパターン辞書を利用し，再現率 98.6%を得ている．本研究では BACT を用いることで部分木を分類の素性とし，異形を吸収しつつ機械学習による曖昧性解消を行う手法を提案する．以下の章では 2 章で慣用句同定手法について述べる．3 章で実験の結果を示し，4 章で結果の考察，5 章でまとめを述べる．

2　慣用句同定手法

本研究では慣用句同定は統計的学習モデルを利用するが，一方で利用方法によっては先行研究 [6][7] で示されているように，可能な慣用句を取りこぼさない高い再現率が求められる場合がある．そこでまず本節では，先行研究 [3] が提案した，慣用句の一部に語が挿入された場合で柔軟にマッチできる入力文グラフについて説明する．次に BACT を利用した慣用句同定手法を提案する．これらの識別結果は後ほど 3 章の実験で比較する．

2.1　入力分グラフ

先行研究 [3] では対象となる文を入力文グラフと呼んでいるグラフに変換する．例文「友人の世話には骨をかなり折っただろう」に対し形態素解析及び係り受け解析を行う．得られた解析結果から，各形態素をノードとして接続し図 2 のグラフを生成する．グラフにおける実線の単線は形態素の並びを，2 重線は係り受け関係を示している．破線の単線は文節を，2 重線は係助詞，接続助詞をスキップしての接続を示している．この構造によって単語の挿入や係助詞，接続助詞の挿入といった慣用句の異形に対応している．また，表層情報のみでなく，原形や読み，助詞の対応といった情報を付与することで表記ゆれや活用，助詞の置換といった慣用句の異形に対応している．

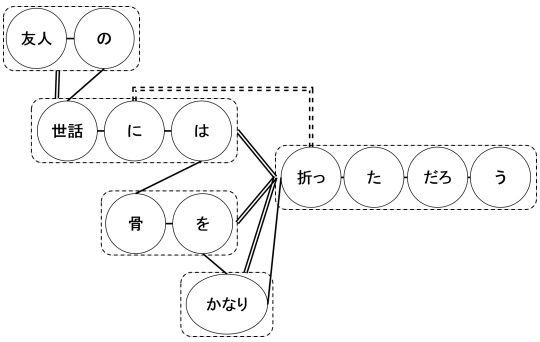

図 2: 入力文グラフ

2.2 BACT

BACT は工藤ら [4] による Boosting アルゴリズムを用いた分類器である．Boosting とは多数の弱学習器を組み合わせることで高精度な学習器を生成する手法である．BACT は構造木を入力として学習させることができ，Boosting に用いる弱学習器に Decision Stumps を用いていることから部分木の有無を素性として分類を行う．例として前節の「友人の世話には骨をかなり折っただろう」であれば,「世話」のような各形態素や係り受け関係，その列などが分類の素性として用いられる．

2.3 素性

本研究では BACT を用いた学習において 5 つのモデルを構築し，比較を行った．各モデルで用いた素性は下記の通りである．

形態素列

MeCab を用いて形態素解析を行ったもの．

係り受け関係

CaboCha を用いて係り受け解析を行ったもの．

名詞カテゴリ

名詞カテゴリ 10 種体系 [5] によって名詞をカテゴリに変換したもの．

品詞列

形態素列を品詞列に置き換えたもの．

慣用句構成文節の情報

慣用句を構成する文節の以下の情報を付与したもの．

態 能動，受動，使役，可能．

法 仮定，命令，願望，疑問，いずれでもない．

文要素 述語，連体修飾，連用修飾．

時制 過去．

極性 肯定，否定．

3 実験

本研究において構築した 5 つのモデルに対し，10 分割交差検定による評価を行い比較した．本章では利用する言語資源及び評価方法について述べた後，各評価結果について報告する．

3.1 評価方法

本実験では，橋本らによって構築された日本語慣用句コーパスから正例，負例が共に 50 以上利用可能である慣用句 93 種を対象とし，各慣用句ごとに評価を行った．

評価は以下の指標を用いて行う．

$$Accuracy = \frac{\text{正しく判定できた用例数}}{\text{全用例数}} \quad (1)$$

$$Precision = \frac{\text{慣用句だと判定した正例数}}{\text{慣用句だと判定した用例数}} \quad (2)$$

$$Recall = \frac{\text{慣用句だと判定した正例数}}{\text{正例数}} \quad (3)$$

各慣用句ごとに用例を 10 グループに分割し，対象 1 グループをテストデータ，残り 9 グループをトレーニングデータとして評価する．これを 10 グループ全てに行い，評価値の平均を算出することで評価結果を得る．

3.2 結果

表 1 に慣用句全体結果及び先行研究の結果を示す．各慣用句ごとの結果は表 2 の通りである．

表 1: 全体の実験結果

モデル	Accuracy	Precision	Recall
形態素	87.87	86.50	86.26
係り受け	87.72	86.24	86.11
名詞カテゴリ	87.28	86.20	85.63
品詞列	84.05	82.08	81.77
慣用句情報	88.05	86.72	86.37
橋本ら [2]	89.19	-	-
守屋ら [3]	72.70	73.20	98.60

表 2: 各慣用句ごとの実験結果 (慣用句情報)

慣用句	Accuracy	Precision	Recall
青筋を立てる	83.72	88.98	92.66
あぐらをかく	91.39	93.56	92.82
足が付く	85.97	74.59	77.02
足が出る	87.70	75.12	68.33
足元を見る	84.57	86.75	86.87
足を洗う	90.55	92.80	93.41
足を伸ばす	93.08	94.99	96.54
頭が痛い	79.15	77.52	73.46
頭を抱える	90.25	93.05	96.07
頭をもたげる	90.21	93.17	95.35
脂が乗る	89.92	73.83	60.21
油を売る	91.45	94.14	96.24
油を絞る	89.40	88.27	79.52
網を張る	80.13	69.52	60.56
息が詰まる	76.28	82.08	85.85
一から十まで	94.25	96.32	97.53
色を失う	85.14	75.24	66.45
腕が上がる	88.08	89.04	90.42
尾を引く	92.60	94.18	97.62
顔を出す	85.92	89.95	93.87
肩を並べる	92.56	95.58	96.20
角が取れる	75.78	76.95	83.24
唇をかむ	78.94	84.48	86.39
口を切る	87.44	87.84	86.67
口をとがらせる	88.27	91.75	95.02
首が回らない	85.05	88.93	88.77
首を切る	83.28	84.45	85.02
首をひねる	95.32	96.55	98.52
事によると	90.05	92.08	93.46
ごまをする	84.51	87.29	81.39
背を向ける	83.14	87.33	87.92
血が通う	76.55	76.89	77.03
力を入れる	95.81	97.20	98.44
宙に浮く	84.63	81.73	82.44
土が付く	87.12	81.73	72.86
手が届く	84.14	88.12	92.96
手がない	91.79	94.49	96.22
手が離れる	90.16	91.95	87.22
手に乗る	90.04	88.79	85.48
手を入れる	88.14	89.81	91.70

表 3: 各慣用句ごとの実験結果 (慣用句情報)

慣用句	Accuracy	Precision	Recall
手を打つ	95.11	96.77	98.14
手を掛ける	90.12	84.32	83.42
手を切る	89.16	89.42	92.50
手を取る	92.72	71.34	63.78
手を握る	94.55	72.43	62.14
手を延ばす	91.84	65.92	50.78
手を広げる	88.18	91.28	92.23
手を回す	90.66	85.17	87.77
峠を越す	86.93	90.43	91.69
泥を塗る	90.26	92.95	94.10
波に乗る	92.61	94.65	96.93
熱が冷める	91.40	93.71	96.96
熱を上げる	94.23	95.87	98.00
熱を入れる	91.47	94.25	95.82
根を下ろす	91.35	93.77	96.36
根を張る	82.21	85.54	84.94
バスに乗り遅れる	92.69	85.45	82.97
バトンを渡す	80.44	84.17	86.41
鼻息が荒い	72.88	75.58	73.92
鼻が高い	83.96	83.88	84.52
鼻を折る	82.17	81.90	77.14
鼻を鳴らす	80.22	82.20	82.83
腹を割る	95.98	97.34	98.49
歯を食い縛る	63.22	71.52	73.16
人を食う	93.68	95.91	95.72
火花を散らす	85.30	88.36	92.96
筆を入れる	81.43	86.92	89.29
船をこぐ	80.82	81.76	81.51
骨が折れる	91.28	92.48	93.69
骨を埋める	91.89	93.78	96.69
骨を折る	90.67	86.07	91.39
幕が開く	84.23	86.34	85.17
右から左	88.93	80.79	75.78
水と油	90.34	91.15	91.51
水に流す	87.41	89.69	91.98
身に付ける	94.08	95.99	97.66
耳が痛い	88.18	85.83	84.98
耳に入れる	86.01	89.72	91.97
実を結ぶ	93.59	95.61	97.34
胸が痛む	94.95	96.00	98.74

表 3: 各慣用句ごとの実験結果 (慣用句情報)

慣用句	Accuracy	Precision	Recall
胸が膨らむ	95.39	96.96	92.73
胸を打つ	96.86	97.66	99.00
目が覚める	91.73	19.17	17.50
芽が出る	86.72	85.68	83.75
目がない	93.02	95.20	97.30
メスを入れる	96.15	96.97	98.78
目に入る	89.57	92.08	96.07
目を覆う	93.91	95.88	97.20
目を覚ます	86.69	65.58	50.23
目をつぶる	87.92	90.29	93.03
目を細くする	70.33	68.29	70.23
指をくわえる	96.35	96.87	99.31
弓を引く	94.20	80.93	68.08

4 考察

前章の実験結果から，本研究で構築した 5 つの素性の比較では，形態素列が 87.87%，係り受け関係が 87.72%，名詞カテゴリが 87.28%，品詞列が 84.05%，慣用句構成文節の各種情報が 88.05%となり慣用句構成文節の各種情報が有意であることが示された．素性ごとの慣用句個別の結果は同様の傾向を示しており，特定の慣用句に対して特定の素性が有意というものは確認できなかった．

また先行研究との比較では，対象慣用句などの条件が同様の橋本らとは全体結果では同等程度の結果となった．慣用句個別では，いくつかの慣用句にては本研究の手法が有意と思われるものがあることが確認できた．守屋らとは条件が違うため単純な比較はできないが，正解率，適合率の両方とも上回った．

今後の課題として，本研究で用いた素性の適用範囲や複数の素性の組み合わせ，新たな素性の利用を考えている．さらに，[6][7] のように慣用句の出現パターンを素性として用いるモデルの構築も考えている．また，実験に用いた日本語慣用句コーパスにはアスタリスクや鍵括弧，空白文字といった記号が多く含まれており BACT の特徴ファイルにも多く出現していた．今回は先行研究と条件を揃えるという点から手を加えなかったが，今後の実験では修正したものを用いての結果も用意する．

5 おわりに

本研究では BACT を用いた日本語慣用句の曖昧性解消を行った．日本語慣用句コーパスを用いた実験では先行研究と同等程度の精度で曖昧性解消を行うことができた．また，素性ごとの比較を行うことで有意な素性を示すことができた．素性の適用範囲や組合せが今後の課題である．

参考文献

[1] 池田 吉優, 竹内 孔一. 意味役割と述語の概念を付与するシステムの構築. 信学技報. Vol. 114, NLC2014-39, pp. 55-60, 2014.

[2] 橋本 力, 河原 大輔. 日本語慣用句コーパスの構築と慣用句曖昧性解消の試み. 情報処理学会研究報告, 2008-NL-186, pp. 1-6, 2008.

[3] 守屋 将人, 竹内 孔一. 網羅的な検出を重視した異形パターンに基づ く日本語慣用句同定システム. 信学技報. Vol. 111, NLC2011-30, pp. 45-50, 2011.

[4] 工藤 拓, 松本 裕治. 部分木を素性とする Decision Stumps と Boosting Algorithm の適用. 電子情報通信学会技術研究報告自然言語処理, 2003-NL-158, pp55-62, 2003.

[5] 森安 祐樹, 竹内 孔一. サ変名詞を含む複合名詞の語義解析システム及び名詞辞書の構築. 信学技報, vol. 111, NLC2011-31, pp. 51-56, 2011.

[6] 竹内 孔一, 白石 貴大 , Ulrich Apel, 宮田 玲, 足立 諒子, Wolfgang Fanderl, 村山 遼, Iris Vogel, 影浦 峡, 簡単なイディオム異形規則の作成: プラットフォームと日本語の異形規則. 言語処理学会第 20 回年次大会発表論文集, pp.488-491, 2014.

[7] 山田 翔平, 矢田 竣太郎, 宮田 玲, 竹内 孔一, Ulrich Apel, Wolfgang Fanderl, 村山 遼, Iris Vogel, 影浦 峡. 日本語イディオム異形規則の構築. 言語処理学会第 21 回年次大会発表論文集, pp. 91-94, 2015.

[8] 佐藤 理史. 基本慣用句五種対照表の作成. 情報処理学会研究報告, 2007-NL-178, pp. 1-6, 2007.