2T-03

動詞出現頻度の偏りを用いた慣用表現の曖昧性解消

舩藤 裕文　　寺岡 丈博
拓殖大学工学部情報工学科

1 背景と目的

慣用句には，慣用表現を考慮した本来の意味と字義的な意味があり，曖昧性を持つものがある．例えば,「骨が折れる」には，体の骨が折れるという字義的な意味と，苦労するという慣用表現としての意味が存在する．文中でこのような表現を検出した上で，それが慣用表現としての意味であるのか，字義的な意味であるのかを正しく判断できなければ，文全体の意味が大きく異なってしまう．以上のことから，慣用表現の曖昧性解消が重要となる．

先行研究 [1][2] では，OpenMWE コーパスのデータを用いて，日本語慣用表現の曖昧性解消が行われている．しかし，これらの手法は「骨が折れる」,「波に乗る」といった慣用句の種類ごとに学習しているため，コーパス外の慣用表現や，新しく慣用的に使われるようになった表現を判定することができない．また，この問題を解決するには，慣用句ごとの用例に対して人手によってアノテーションを行い，OpenMWE コーパスを拡充するほかなく，膨大な種類がある慣用表現に対応することは難しい．

そこで，本研究では OpenMWE コーパスに依存しない日本語慣用表現の曖昧性解消の精度向上を目的とする．

2 提案手法

本研究では，慣用表現が含まれる文（イディオム）に出現する語と，慣用表現を含まず，字義通りにしか解釈できない文（リテラル）に出現する語の出現頻度の違いから判断する手法を提案する．

2.1 使用する言語資源

言語資源として OpenMWE コーパス [1] と新聞コーパスを用いる．新聞コーパスには CD-毎日新聞データ集の 2017 年版と 2018 年版を使用する．

OpenMWE コーパスとは，慣用句の意味について用例ごとにイディオムであるか，リテラルであるかを人手によってアノテーションされたコーパスである．収録された慣用句の種類は 135 句であり，全体の用例数は 102,334 文である．

2.2 処理の流れ

処理の流れを図 1 に示す．本手法は学習フェーズと曖昧性解消フェーズの二段階からなる．学習フェーズは，OpenMWE コーパスの用例から決定木モデルを学習させるフェーズである．OpenMWE コーパスの一文に対して形態素解析，構文解析を行う．その結果から語の出現頻度などの学習に使用する素性を計算する．決定木の学習に使用した素性を表 1 に示す．ここで，表 1 における n_1，n_2，v とは，それぞれ，n_1 が文中の一つ目の名詞，n_2 が文中の二つ目の名詞，v が文中の動詞を表す．また，本研究では，n_1 は，n_2，v のどちらかに係り関係を持ち，n_2 は，v に係り関係を持つ文を対象とする．素性の計算には新聞コーパスを使用しており，表 1 の文書数とは，新聞コーパスの記事を単文ごとに分けた文書数である．曖昧性解消フェーズは，学習フェーズで学習させた決定木モデルを用いて，入力文がイディオムかリテラルかを二値で出力するフェーズである．

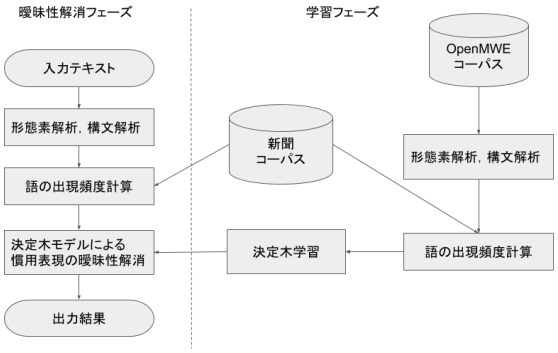

図 1: 処理の流れ

3 評価実験

3.1 実験方法

本研究の評価として，OpenMWE コーパスの用例に対して決定木モデルを構築し，10 分割交差検証を行った．実験データには，OpenMWE コーパスに収録されている慣用句のうち，本研究の対象である，名詞，助詞，動詞の三語で構成される慣用句 106 句の用例の中

Disambiguation of Japanese Idiomatic Expression using Difference of Verb Frequency
Hirofumi Funato and Takehiro Teraoka
Department of Computer Science, Faculty of Engineering, Takushoku University

表 1: 学習に使用した素性一覧

素性	説明
F1	n_1,n_2 が共起する文書数 / n_1 が含まれる文書数
F2	n_1,n_2 が共起する文書数 / n_2 が含まれる文書数
F3	n_1,v が共起する文書数 / n_1 が含まれる文書数
F4	n_1,v が共起する文書数 / v が含まれる文書数
F5	n_2,v が共起する文書数 / n_2 が含まれる文書数
F6	n_2,v が共起する文書数 / v が含まれる文書数
F7	n_1 に共起する動詞の異なり数 / n_1 が含まれる文書数
F8	n_1 に共起する名詞の異なり数 / n_1 が含まれる文書数
F9	n_2 に共起する動詞の異なり数 / n_2 が含まれる文書数
F10	n_2 に共起する名詞の異なり数 / n_2 が含まれる文書数
F11	v に共起する名詞の異なり数 / v が含まれる文書数
F12	n_1,n_2 に共起する動詞の異なり数 / n_1 に共起する動詞の異なり数
F13	n_1,n_2 に共起する動詞の異なり数 / n_2 に共起する動詞の異なり数
F14	n_1,v に共起する名詞の異なり数 / n_1 に共起する名詞の異なり数
F15	n_1,v に共起する名詞の異なり数 / v に共起する名詞の異なり数
F16	n_2,v に共起する名詞の異なり数 / n_2 に共起する名詞の異なり数
F17	n_2,v に共起する名詞の異なり数 / v に共起する名詞の異なり数

から，名詞を二つ，動詞を一つ含む用例を抽出し，イディオムの用例を 500 文，リテラルの用例 500 文の合計 1000 文を使用した．

先行研究 [1] では，評価対象の全用例をイディオムかリテラルの多い方に一律に解釈したものをベースラインとしている．本研究も，先行研究に倣い，全ての用例を一律にイディオムとして解釈したものをベースラインとした．

3.2 結果

表 2 は，評価データ全体の正解率，イディオムの検出における適合率，再現率，そして F 値についてまとめたものである．提案手法は再現率以外の項目でベースラインを上回った．また，ベースラインのみが正解した用例と，提案手法のみが正解した用例を用いて，符号検定を行った結果，有意差が得られた（p<0.01）．

表 2: 評価実験の結果．*はベースラインに対する統計的な有意差を示す．(*p<0.01)

手法	正解率	適合率	再現率	F 値
ベースライン	0.50	0.50	1.00	0.67
提案手法	0.73*	0.73	0.72	0.72

3.3 考察

評価データの全用例を学習させた決定木モデルにおける素性ごとの重要度を図 2 に示す．重要度とは，決定木において一つの素性が影響を与えたノードが多いほど値が高くなる指標である．図 2 において最も重要度が高かった素性である F12 について考察を述べる．

F12 は，n_1，n_2 に共起する動詞の異なり数を n_1 に共起する動詞の異なり数で割った値である．これは値が小さいほどイディオムであり，値が大きいほどリテラルである．つまり，n_1 に共起する動詞の種類に比べて，n_1 と n_2 に共起する動詞の種類が少ないほどイディオムであり，n_1 と n_2 に共起する動詞の種類が多いほどリテラルであるということが判断できる．

また，本研究の性質として，新聞コーパスに記載されていない慣用表現の判断ができないことが挙げられる．しかし，この問題は，新聞記事や，web 上の文書がまとめられたコーパスを新たに追加することで解決できる．

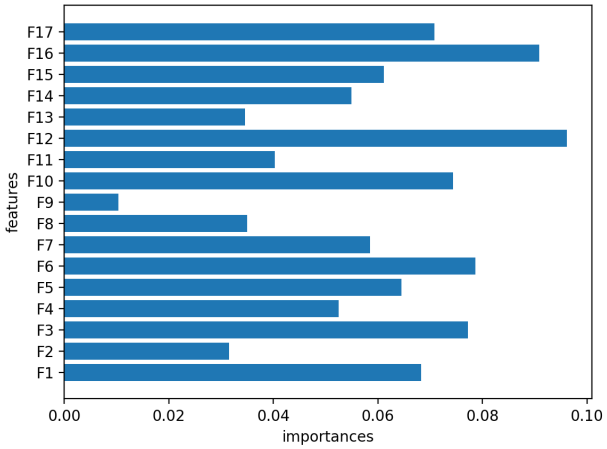

図 2: 素性ごとの重要度

4 まとめ

本研究は，入力文がイディオムかリテラルかを入力文に含まれる名詞，動詞の出現頻度に基づいて判定する手法を提案した．本手法は語の出現頻度という慣用句の種類に左右されない情報を素性として判断を行っているため，本研究の対象である名詞，助詞，動詞の三語から構成される慣用表現であれば，OpenMWE コーパスに収録されていない慣用表現が含まれる文でも判断できることが強みである．

今後の課題として，本研究で対象とした構成以外の形をとる慣用表現に対応することなどが考えられる．

謝辞

本研究の一部は JSPS 科研費 18K12434 の助成を受けたものである．

参考文献

[1] 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, 2017.

[2] 宮田周, 竹内弘一. 統計的学習モデルを利用した日本語慣用句の意味的曖昧性解消. 情報処理学会第 79 回全国大会講演論文集（1），pp.599-600, 2017.