自動検出のための慣用句の分類と語彙的情報

橋本 力[*] 佐藤 理史[†] 宇津呂 武仁[‡]

[*‡]京都大学大学院情報学研究科 [†]名古屋大学大学院工学研究科

{[*] hasimoto,[‡] utsuro}@pine.kuee.kyoto-u.ac.jp [†] ssato@nuee.nagoya-u.ac.jp

1 はじめに

慣用句検出技術は正確な言語理解に欠かせない。慣用句検出の失敗は、例えば機械翻訳の失敗につながる。Excite の翻訳サイト[1]では、慣用句「骨を折る」の検出に失敗するため、(1a) を (1b) のように誤訳してしまう。

(1) a. 彼は問題の解決に骨を折った。
b. He broke his bone to the resolution of a question.

慣用句検出のためには慣用句辞書と検出器が必要である。しかし、現在までに、広く利用可能な慣用句辞書と検出器は開発されていない。

本研究では、慣用句辞書の構築に向けて、検出に必要な慣用句の分類と語彙的情報について考察する。本研究で言う慣用句検出とは、慣用句の字面を見つけるだけでなく、その字面が実際に慣用句の意味で用いられているかどうかまで識別する処理を言う。(1a) の場合、「骨を折っ」を慣用句として検出するが、「足の骨を折った」の場合は検出しない。[2]

なお、本研究では、慣用句を「複数文節から成る、意味的に非構成的な表現」と定義しておく。

2 検出のための慣用句の分類

2.1 慣用句検出の難しさ

慣用句検出の難しさには 2 種類ある。

(2) a. 出現形態の変化に起因するもの
b. 慣用句の意味と文字通りの意味との間の曖昧性に起因するもの

(1b) の検出誤りは慣用句の曖昧性に起因するものである。出現形態変化に起因する誤りの例として (4b) を挙げる。

(3) a. 彼は役に立たない。
b. He is useless.

(4) a. 彼は役には立たない。
b. He doesn't stand in the post.

(3b)、(4b) ともに Excite による翻訳である。慣用句「役に立つ」を含む文 (3a) は正確に訳せるのに、その慣用句に助詞「は」が挿入されただけの文 (4b) は誤訳してしまう。

2.2 検出のための慣用句の分類

2.2.1 検出難易度による分類

検出に必要な語彙的情報は、検出の難易度によって変わる。そこで、慣用句を検出難易度によって分類する (図 1 左)。慣用句の検出難易度は、(2a) と (2b) の 2 つの観点により決まる。

クラス A: 形態変化不可能で、曖昧性もない。
クラス B: 形態変化は可能だが、曖昧性はない。
クラス C: 形態変化が可能で、曖昧性もある。

クラス C の慣用句の検出が最も難しい。クラス A の慣用句は曖昧性の無い一単語に相当し、検出は容易である。

2.2.2 相当品詞と内部構造による分類

クラス C の慣用句の検出に必要な語彙的情報は、相当品詞と内部構造によって変わる。一方、クラス A、B の場合は変わらない。内部構造は、慣用句構成語の品詞と、構成語間の依存関係によって決まる。

クラス C の慣用句を検出するには、慣用句の意味と文字通りの意味との間の曖昧性を解消する必要がある。そのためには、慣用句として用いられる場合と、文字通りの意味で用いられる場合との間の、用法上の違いに注目する必要があ

[1] http://www.excite.co.jp/world/

[2] ある慣用句は、複数の慣用句としての意味を持つ。例えば「顔を見せる」は、「特徴を示す」という慣用的意味と (「保守派の顔を見せる」)、「出席する」という慣用的意味を持つ (「同窓会に顔を見せる」)。本研究で言う検出は、複数の慣用的意味の区別までは含まない。

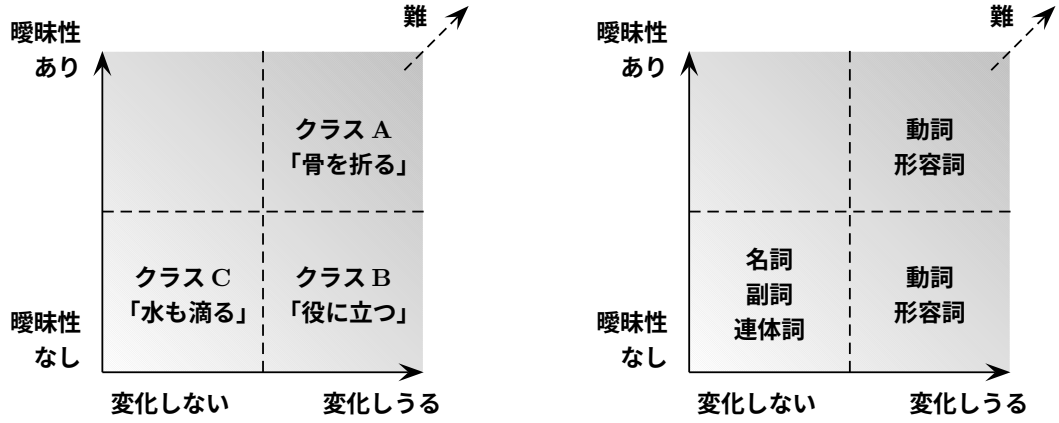

図 1: 検出難易度による分類（左）と相当品詞ごとの分布（右）

る。曖昧性解消に用いる、用法上の違いに関する情報を、曖昧性解消情報と呼ぶ。例えば「骨を折る」は、受け身の場合 (5a) や分離した場合 (5b) は、文字通りの意味しかない。

(5) a. 骨が折られる
b. 骨を何度も折る

曖昧性解消情報は、相当品詞と内部構造に応じて変わる。

相当品詞： 動詞相当句の場合には、形容詞相当句の場合と異なり、例えば、受け身形の有無が曖昧性解消情報の一つになる。

内部構造： 慣用句構成文節の分離の有無も曖昧性解消情報の一つになる。分離の有無を調べるべき文節境界の数は、3 文節から成る慣用句は 2 箇所だが、2 文節のものは 1 箇所である。

クラス A、B は、クラス C と違い、細分化する必要はない。クラス A の慣用句の検出には表記のみで十分である。表記情報に相当品詞あるいは内部構造の違いは関係ない。クラス B の慣用句に必要なのは、出現形態の変化を検出時に吸収するための語彙的情報である。本研究では、慣用句構成語間の依存関係をこの語彙的情報として用いる。これを依存構造情報と呼ぶ。この情報も相当品詞または内部構造の違いに左右されない。

つまり、クラス C のみが相当品詞と内部構造による細分類が必要である。また、検出難易度による分類と慣用句相当品詞の間には、図 1 右のような相関がある。

以上を考慮すると、検出のための慣用句の分類体系は図 2 のようになる。内部構造の全種類の把握は今後の課題である。

3 検出に必要な語彙的情報

3.1 クラス別の語彙的情報

クラス別の語彙的情報は、表 1 のようにまとめられる。

表 1: クラス別の語彙的情報

	表記情報	依存構造情報	曖昧性解消情報
クラス A	○		
クラス B	○	○	
クラス C	○	○	○

3.1.1 クラス A

クラス A の慣用句は曖昧性の無い一単語に相当するので、検出に必要な情報はその表記だけである。ただし、以下の 3 種類の表記の揺れを吸収する必要がある。

- 漢字, 平仮名の違い:「{ 怨み/うらみ } を買う」
- 送り仮名の有無:「{ 折り紙/折紙 } をつける」
- 接頭辞「お」の有無 :「{ お役/役 } に立つ」

3.1.2 クラス B

クラス B の慣用句には、表記だけでなく、以下の 3 種類の出現形態変化に対応するための語彙的情報が必要である。

- 構成要素の分離 ..... 「役にすごく立つ」
- 述語の変化 ..... 「役に立てば」
- 助詞の変化 ..... 「役には立つ」

本研究では、この語彙的情報として、慣用句構成語間の依存関係 (依存構造情報) を用いる。ただし、述語の活用語尾と助詞「が」「を」は、変化または消失しうるので無視する。

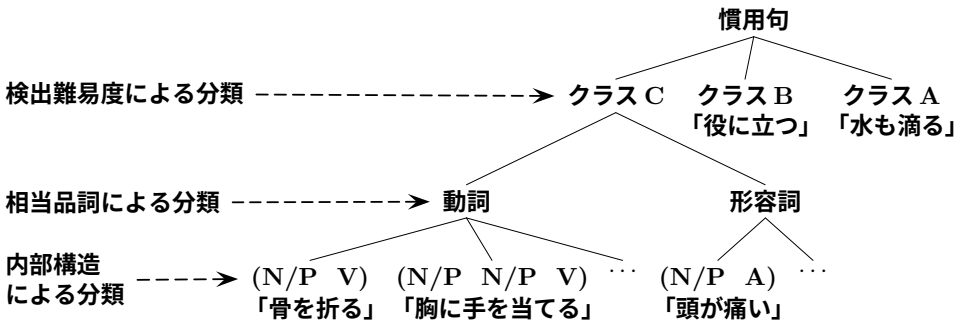

図 2: 検出のための慣用句の分類体系

3.1.3 クラス C

クラス C には、表記情報と依存構造情報に加えて、慣用句の意味と文字通りの意味との間の曖昧性を解消するために、曖昧性解消情報が必要である。曖昧性解消情報として、慣用句として用いられる場合と、文字通りの意味の句として用いられる場合との間の、用法上の差異を利用する。通常、慣用句としての用法の方が、文字通りの意味の句の用法より、文法上の制約が厳しい。曖昧性解消情報として利用できるのは、この、文字通りの句の場合には可能だが、慣用句の場合は不可能な用法である。

§2.2.2 で述べた通り、曖昧性解消情報として何が利用できるのかは、相当品詞と内部構造に応じて異なる。検出のための慣用句辞書を構築するには、相当品詞と内部構造による、クラス C の全ての下位分類に対して、利用可能な曖昧性解消情報を明らかにする必要がある。

以下では、第一段階として、(N/P V) 型動詞慣用句の曖昧性解消情報を考察する。(N/P V) 型動詞慣用句は、「骨を折る」のように、名詞、助詞、動詞から構成される、動詞相当の慣用句である。

3.2 (N/P V) 型動詞慣用句の語彙的情報

3.2.1 (N/P V) 型動詞慣用句

(N/P V) 型は、タイプ数、トークン数ともに、慣用句中で最多である。よって、本研究では、この種類から詳細に扱うことにした。

タイプ数に関して、学研国語大辞典 (金田一・池田, 1989) 慣用句見出し 4,581 句のうち 1,834 句、つまり 40%が (N/P V) 型であった。[3]

トークン数に関して、学研慣用句 4,581 句と毎日新聞 10 年分 ('91–'00) を用いて、慣用句頻度調査を行った。[4] 毎日新聞 10 年分の全慣用句トークンは 220,684 句であった。そのうち (N/P V) 型は 167,268 句、つまり 75.79%を占めた。

3.2.2 (N/P V) 型動詞慣用句の語彙的情報

宮地 (1982, 1985) や森田 (1985) などの日本語学における慣用句研究で、慣用句の場合と、文字通りの意味の句の場合との間の用法上の差異がまとめられている。このうち、(N/P V) 型動詞慣用句の曖昧性解消に利用できそうなものを選定した。

(6) (N/P V) 型動詞慣用句の曖昧性解消情報

a. N を連体修飾することができる文法範疇
   I. 関係節
   II. 属格句
   III. 連体詞

b. P を置換、あるいは P に付加することができる提題・取り立て助詞
   I. 「は」「も」
   II. 「は」「も」以外の提題・取り立て助詞

c. V に付加できる助動詞・接尾動詞
   I. モダリティ
      i. 肯定・否定形
      ii. 意志動詞が取りうるモダリティ形式 [5]
   II. ヴォイス
      i. 受け身

[3] 慣用句見出しを茶筌 (松本他, 2000) で形態素解析し、品詞列ごとに見出し数を集計した。集計結果の人手によるチェックは行っていない。全慣用句見出しは 4,802 句だが、そこから茶筌未知語を含む慣用句を除いた 4,581 句を対象とした。

[4] 頻度計算は文字列照合により行った。ただし、動詞部分の屈折に対応するために、あらかじめ、次のことをした：慣用句見出しと新聞記事を、ともに、形態素解析器により、基本形の列に変換しておく。この頻度計算も全て自動で行っており、人手によるチェックは行っていない。

[5] 具体的には次の形式を指す：命令、禁止、許可、意志、依頼。表現のリストは益岡・田窪 (1992) から得た。

表 2: 検出のための慣用句辞書の構成

ID	クラス	表記	依存構造	下位分類
011	A	水も滴る	–	–
		...		
022	B	役に立つ	(役/に 立つ)	–
		...		
033	C	骨を折る	(骨/を 折る)	(N/P V)

(N/P V) 動詞句用曖昧性解消情報

	連体	提題	受身	分離	...
...			...		
骨を折る	×	○	×	×	...
...			...		

ii. 使役

d. (N/P V) 全体の属性

I. 慣用句構成文節の分離

II. 項の選択制限

例えば、「骨を折る」が慣用句として使われる場合、連体修飾 (6a) は一切受けつけない。(7b) は属格句による連体修飾の例だが、(7a) と違い、文字通りの意味にしか解釈できない。

(7) a. 彼に骨を折らせてしまった。
b. 彼の骨を折らせてしまった。

4 辞書の構成

以上の知見をもとに慣用句辞書を構築した。辞書の構成は、図 2 の分類体系に従っている (表 2)。まず、全ての慣用句に ID、所属クラス、表記情報が与えられる。クラス B、C にはさらに依存構造情報が与えられる。最後に、クラス C にのみ、曖昧性解消情報が与えれる。§3.1.3 で述べた通り、曖昧性解消情報は下位分類によって異なる。そこで、下位分類ごとの曖昧性解消情報テーブルを、辞書とは別に用意した。辞書のクラス C エントリには、該当する曖昧性解消情報テーブルへのポインタが与えられる。

5 関連研究

本研究では、従来無かった、検出難易度に基づく分類を提案した。

宮地 (1982, 1985) などの日本語学における研究では、慣用句の様々な分類が提案されてきた。自然言語処理においても分類がいくつか提案された (奥, 1990; Shudo et al., 2004)。それらの多くは、意味的構成性と統語的な固定性に基づくものである。

本研究でも、検出難易度を決定する軸の 1 つとして統語的固定性に相当するものを利用している。しかし、意味的性質として、検出のためには、構成性よりも曖昧性の有無の方が重要である。

6 おわりに

本研究では、慣用句を自動検出するのに必要な分類と語彙的情報について考察した。自動検出のためには、慣用句を、検出難易度、相当品詞、内部構造の観点で分類すべきことを論じた。検出難易度別に、検出に必要な語彙的情報について考察した (表 1)。クラス C の曖昧性解消情報について、(N/P V) 型動詞慣用句を取り上げて論じた (6)。また、本研究の慣用句辞書の構成について述べた。

今後の課題は次の 2 点である。i) クラス C の慣用句の内部構造を全て把握する。ii) クラス C の下位分類全てについて、曖昧性解消情報を明らかにする。

謝辞 本研究のために、学研国語大辞典の慣用句見出しの利用を許諾してくださった学習研究社に感謝申し上げます。

参考文献

Shudo, K., Tanabe, T., Takahashi, M., & Yoshimura, K. (2004). MWEs as Non-propositional Content Indicators. In *the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 32–39.

奥雅博 (1990). 「日本語解析における述語相当の慣用的表現の扱い」. 『情報処理学会論文誌』, **31** (12), 1727–1734.

金田一春彦・池田弥三郎 (編) (1989). 『学研国語大辞典第二版』. 学習研究社.

宮地裕 (1982). 『慣用句の意味と用法』. 明治書院.

宮地裕 (1985). 「慣用句の周辺 —連語・ことわざ・複合語—」. 『日本語学』, **4** (1), 62–75.

松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 高岡一馬, 浅原正幸 (2000). 『日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書』. 奈良先端科学技術大学院大学.

益岡隆志・田窪行則 (1992). 『基礎日本語文法 改訂版』. くろしお出版.

森田良行 (1985). 「動詞慣用句」. 『日本語学』, **4** (1), 37–44.