依存構造照合に基づく慣用句自動検出

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、慣用句の検出に失敗するため、(1b) のような誤訳をしてしまう。

(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。

本研究では検出器を構築する。検出器は、橋本他 (2006) で提案した語彙的情報を利用する。

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 検出のための分類と語彙的情報

2.1 検出のための慣用句の分類

慣用句検出の難しさは、i) 出現形態変化の有無と、ii) 慣用句の意味と文字通りの意味との間の曖昧性の有無によって決まる。慣用句はその検出難易度によって、A、B、C の 3 クラスに分類される。クラス A は曖昧性の無い一単語に相当し、検出は用意である。クラス C の検出には、出現形態変化と曖昧性に対応せねばならず、最も難しい。本研究では、クラス B と C を対象にする。

クラス A: 形態変化不可能で、曖昧性もない。
(「水も滴る」)
クラス B: 形態変化は可能だが、曖昧性はない。
(「役に立つ」)
クラス C: 形態変化が可能で、曖昧性もある。
(「骨を折る」)

2.2 クラス B の語彙的情報

クラス B には、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。その情報として、慣用句構成語間の依存関係を用いる。これを依存構造情報と呼ぶ。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。

2.3 クラス C の語彙的情報

クラス C には、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。

利用できる用法上の差異は、慣用句の相当品詞と内部構造に応じて異なる。橋本他 (2006) では、使用頻度が最も高い (N/P V) 型動詞慣用句を対象に、(2) に挙げた曖昧性解消情報を提案した。

(2) **(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」

[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。

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[3]

II. ヴォイス

i. 受け身

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

3 依存構造照合に基づく検出

依存構造照合に基づく慣用句検出方法を提案する。依存構造照合とは、慣用句の依存構造が入力文の依存構造に包含されているかを調べる処理である (図 1)。以下に処理手順を示す。

1. 入力文を形態素解析、依存構造解析する。
2. 入力文の依存構造を、辞書にある全ての慣用句の依存構造パターンと照合する。
3. 慣用句が入力文に含まれていたら、該当する慣用句の ID (図 1 では 022) を、以下の箇所に付与して出力する。
   - 慣用句構成語にあたる形態素
   - その形態素を含む文節（図 1 の ）

クラス B と C の扱いの違いは、辞書中の依存構造パターンのみである。クラス B の依存構造パターンは、依存構造情報のみを表す。一方クラス C の依存構造パターンには、依存構造情報に加えて、曖昧性解消情報も盛り込まれる (図 2)。

4 検出器と辞書の構築

4.1 検出器の構築

依存構造照合に基づく検出器を構築した (図 3)。形態素解析は茶筌 (松本他, 2000)、依存構造解析は南瓜 (Kudo & Matsumoto, 2002) によって行われる。依存構造照合は TGrep2 (Rohde, 2005) によって行われる。

依存構造パターンは、TGrep2 の構文パターン検索用言語の仕様に基づいており、複雑で、人間にとって理解しにくい。そこで、人間に理解しやすい形式の辞書から、TGrep2 のパターン集合を自動生成するシステムも併せて構築した (図 3 の依存構造パターン生成器)。

4.2 慣用句辞書の構築

(N/P V) 型動詞慣用句 100 句からなる慣用句辞書を構築した。辞書の構成は橋本他 (2006) の提案に従う。[4]

収録した 100 句は、クラス B または C に属する。100 句の選定にあたっては、日常的によく用いられる慣用句を集めるように配慮した。具体的には、宮地 (1982) の慣用句リストと、金田一・池田 (1989) 中の慣用句から、次の手順で 50 句ずつ選定した。[5]

1. 宮地 (1982) から、毎日新聞 ('91–'00) における最頻出 50 句を抜粋
2. 同様に、金田一・池田 (1989) から最頻出 50 句を抜粋 (ただし、宮地 (1982) から抜き出した 50 句と重複するものは含めない。)

100 句の内訳は、クラス B に属するものが 66 句、クラス C に属するものが 34 句である。

5 検出器の評価実験

5.1 実験の概要

語彙的情報の効果を評価するため、検出器の評価用データに対する性能を調べる実験を行った。

評価用データとして、毎日新聞'95 から、各慣用句につき 3 文ずつ用例を収集した。ただし表 1 の 3 句については、その異表記の用例も収集した。異表記の出現頻度が代表表記の頻度と同様に高かったためである。代表表記 100 句と異表記 3 句のそれぞれに対して用例を 3 文収集したので、合計 309 文となる。309 文の内訳を表 2 に挙げる。左表の値は、該当する用例の数と、全体に占めるその割合である。右表は、クラスごとの用例数の割合を示す。

曖昧性解消情報の効果を検証するために、ベースラインシステムを作成し、その評価も行う。ベースラインシステムは、本研究の検出器と基本的に同じである。違いは、ベースラインシステムは曖昧性解消情報を全く参照しないという点である。

[3] 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。

[4] クラス C の慣用句への曖昧性解消情報の付与に関して、(2) のうち、慣用句の項に対する選択制限 (2dII) の付与は行っていない。

[5] 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。この頻度計算は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。

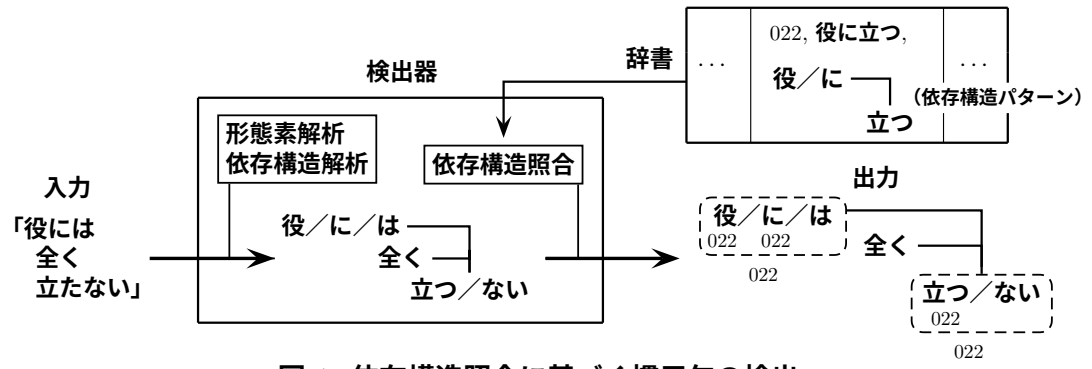

図 1: 依存構造照合に基づく慣用句の検出

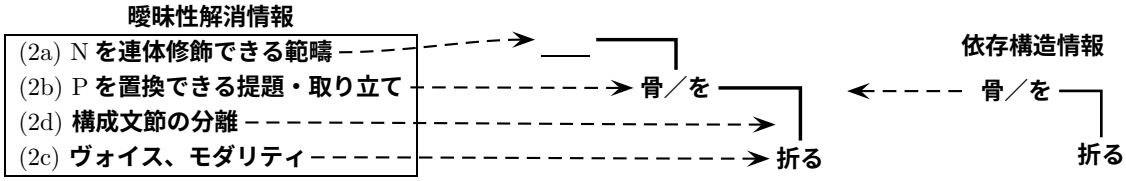

図 2: クラス C の依存構造パターン

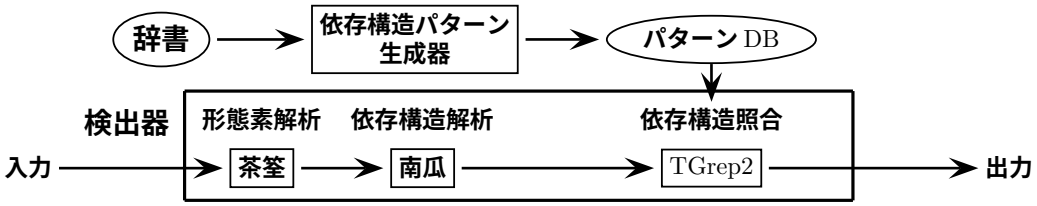

図 3: 検出システムの構成

表 1: 評価用データに用例を加えた異表記

代表表記（頻度, クラス）	異表記（頻度, クラス）
声をかける（4365, C）	声を掛ける（1464, C）
身につける（2491, C）	身に付ける（1277, C）
気がつく（2424, B）	気が付く（860, B）

表 2: 評価用データの概要

	クラス B	クラス C	全データ	クラス B	クラス C
正例	200 (64.72%)	66 (21.36%)	266 (86.08%)	200 (99.50%)	66 (61.11%)
負例	1 (0.32%)	42 (13.59%)	43 (13.92%)	1 (0.50%)	42 (38.89%)
合計	201 (65.05%)	108 (34.95%)	309 (100.00%)	201 (100.00%)	108 (100.00%)

表 3: 検出器（左）とベースラインシステム（右）の検出性能

	クラス B	クラス C	全ての用例
再現率	0.975 ($\frac{195}{200}$)	0.939 ($\frac{62}{66}$)	0.966 ($\frac{257}{266}$)
精度	1.000 ($\frac{195}{195}$)	**0.697** ($\frac{62}{89}$)	**0.905** ($\frac{257}{284}$)
F 値	0.987	**0.800**	**0.935**

	クラス B	クラス C	全ての用例
再現率	0.975 ($\frac{195}{200}$)	0.939 ($\frac{62}{66}$)	0.966 ($\frac{257}{266}$)
精度	1.000 ($\frac{195}{195}$)	0.602 ($\frac{62}{103}$)	0.862 ($\frac{257}{298}$)
F 値	0.987	0.734	0.911

5.2 実験結果

表 3 に実験結果を示す。再現率、精度、F 値の式は以下の通りである。

$$\text{再現率} = \frac{\text{正しく検出された慣用句の数}}{\text{正解慣用句数}}$$

$$\text{精度} = \frac{\text{正しく検出された慣用句の数}}{\text{検出された全ての慣用句の数}}$$

$$\text{F 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}}$$

検出器とベースラインシステムとの違いは太字で示されている箇所である。つまり、検出器の方が検出誤りが少ない。これは、ベースラインシステムと違い、検出器が曖昧性解消情報も参照しているためである。

5.3 考察

(N/P V) 型動詞慣用句は、90%以上の精度で、90%以上検出できることがわかった。ただし、これは、検出器の以下の性質によると考えられる。

- 評価用データ中に負例が少なければ少ないほど、性能が高く見積もられる。

検出器は、クラス C 慣用句候補の表出形態に、100%(近い確率で) 負例と言い切れる特徴がある場合にのみ、負例として排除する。結果として、検出器は、多くの慣用句候補を正例として判断する。そのため、評価実験の結果は、クラス C の負例の評価用データ全体に占める割合に依存すると言える。つまり、負例が少なければ少ないほど、検出性能は高く見積もられる。

実際、クラス C の慣用句に対象を限定すると、90%以上検出できるが、精度は 70%弱に落ちる。また、クラス C の負例 42 文のうち、排除に成功したのは 15 文のみであり、成功率はわずか 35.71% である。

今回収集した評価用データの内訳 (表 2) にあるように、クラス C の負例の使用頻度はそれほど高くないと思われる。しかし、検出精度の向上のためには、正確な負例排除に向けて、曖昧性解消情報を改良していく必要がある。

5.3.1 曖昧性解消情報の効果

クラス C の負例排除に成功した 15 例のうち、負例排除に貢献した語彙的情報は以下の 4 つであった。[6]

1. 属格句による連体修飾 (2aII) ..... 6 例
2. 関係節による連体修飾 (2aI) ..... 5 例
3. 構成文節の分離 (2dI) ..... 2 例
4. 否定形 (2dI) ..... 1 例

負例排除に失敗した 27 例のうち、5 例は、項の選択制限を追加することで排除できると思われるものであった。残り 22 例は、現在の言語処理技術では、排除が困難と思われるものであった。

6 おわりに

本研究では、依存構造照合に基づく慣用句検出の枠組を提案した。また、橋本他 (2006) で提案した、検出のための語彙的情報の効果を、検出実験により評価した。

実験結果から、(N/P V) 型動詞慣用句は、90% 以上の精度で、90%以上検出できることがわかった。しかし、クラス C の慣用句に対象を限定すると、精度は 70%弱に落ちる。また、クラス C の負例排除の成功率はわずか 35.71%である。さらなる精度向上に向けて、曖昧性解消情報の洗練が必要である。

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

Kudo, T. & Matsumoto, Y. (2002). Japanese Dependency Analyisis using Cascaded Chunking. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pp. 63–69 Taipei.

Rohde, D. L. T. (2005). *TGrep2 User Manual version 1.15*. Massachusetts Institute of Technology. http://tedlab.mit.edu/~dr/Tgrep2/.

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.

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.

橋本力, 佐藤理史, 宇津呂武仁 (2006). 「自動検出のための慣用句の分類と語彙的情報」. 『言語処理学会第 12 回年次大会』.

[6] 以下に加えて、南瓜の解析ミスにより負例が排除された例が 1 つあった。