慣用句格フレームの自動構築に向けて*

橋本 力　　河原 大輔
山形大学大学院理工学研究科　情報通信研究機構
ch@yz.yamagata-u.ac.jp　　dk@nict.go.jp

1　はじめに

格フレームは構文解析や意味解析等で重要な役割を果たす。しかし、従来の格フレーム構築研究のほとんどは一語から成る単純述語にのみ焦点を当てており、複合述語は等閑視されてきた。

本研究では、複合述語の中でも使用頻度の高い慣用句を対象とした、コーパス中の用例を用いた格フレーム自動構築法を開発する。我々が構築しようとしている慣用句格フレームは、Kawahara and Kurohashi (2006) 流の、格スロットごとに、そこに入るべき単語 (格インスタンス) を列挙するスタイルのものである。例えば、慣用句「骨を折る」の場合、(1) のようになる。

(1)

委員会 巨人 父親 学生	ガ	収拾 (首位) 奪還 (息子の) ため (暗記する) の	ニ　骨を折る

技術的課題は、格フレームの元となる慣用句用例を収集する際、字面は同じだが文字通りの意味で使われている句（リテラル句）の用例を正確に排除することである。例えば、(2i) は慣用句「骨を折る」の用例として適切だが、(2l) は慣用句用例から排除したい。

(2) i. 委員会が事態の収拾に骨を折った。
　　l. 選手が足の骨を折った。

我々は、慣用句とリテラル句の文法的制約の差異に注目して、慣用句用例のみを正確に収集することを検討している。例えば「骨を折る」が慣用句である場合、その名詞構成素が連体修飾を受けることはない。実際 (2l) では「骨」が連体修飾されており、慣用句としては解釈できず、文字通りの意味しかない。

構築された格フレームは、慣用句検出タスクを通じて評価する予定である。

*本研究の一部は、日本学術振興会科学研究費補助金若手研究 (B)「日本語慣用句の検出と格解析のための言語資源の構築」(課題番号 19700141、研究代表者 橋本力) の援助を得てなされた。

2　要素技術

本研究では、用例集からの格フレーム自動構築の枠組として Kawahara and Kurohashi (2006) (以下、河原ら) を (§2.1)、慣用句用例収集の手法として Hashimoto et al. (2006) (以下、橋本ら) を採用する (§2.2)。

2.1　格フレーム自動構築

格フレーム自動構築では、述語の意味的曖昧性にどう対応するかが問題となる。例えば、「荷物を積む」と「経験を積む」の「積む」は意味が異なると考えられるが、このような違いを適切に反映した格フレームを如何に自動で構築するかが課題である。

河原らは、述語の直前の格に着目して用例をクラスタリングすることで、述語の意味ごとに格フレームを自動構築することに成功した。手法の大枠は次の通りである。

1. コーパスを構文解析し、述語項構造を抽出する
2. 述語とその直前格のペアの同一性に基づいて述語項構造をクラスタリングし、個々のクラスタを 1 次格フレームとする
3. 格フレーム間類似度に基づき 1 次格フレームをクラスタリングし、最終的な格フレーム群を得る[1]

この結果、「荷物を積む」と「積荷を積む」が属する格フレームと、「経験を積む」と「修行を積む」が属する格フレームを区別することが可能となる。つまり、述語の直前格の違いで用法の違いを近似している。

一方、曖昧性のある慣用句の格フレーム構築では、あらかじめ 1 つの慣用句につきクラスタ数が 2 つ (慣用句とリテラル句) と決まっている。また、(1) にあるように、同じ格スロットに入る単語の意味が、河原らの手法が作り上げる格フレームと比べて、より広範囲に渡る。そのため、本研究の慣用句格フレーム構築で

[1]格フレーム間類似度については Kawahara and Kurohashi (2006) を参照。

は、述語 (慣用句) と直前格のペアによるクラスタリングを通して用法 (慣用句あるいはリテラル句) の違いを表すより、あらかじめ用例を慣用句とリテラル句に分けておいた方が良い結果が期待できる。

2.2 慣用句検出

橋本らは、入力文中の慣用句を検出するタスクに取り組んだ。このタスクには、慣用句とリテラル句の曖昧性解消も含まれている。従って、字面だけで慣用句かどうか判断するわけではない。

橋本らは、まず、形態変化の有無と曖昧性の有無に基づいて慣用句を A、B、C の 3 つに分類した。クラス A は形態変化も曖昧性もない最も検出が容易なクラスであり、「水も滴る」などが含まれる。クラス B は曖昧性はないが形態変化はするクラスで、「役に立つ」などが該当する。クラス C は形態変化も曖昧性もある最も検出が難しいクラスで、「骨を折る」などが含まれる。

クラス A の検出には、その字面だけを用意しておけばよい。クラス B の検出には、字面と構成素間の依存関係が分かれば事足りる。一方クラス C の検出には、字面と依存関係に加えて、慣用句と字面が対応するリテラル句を区別するための手掛かりが必要である。

橋本らは、慣用句とリテラル句との間の文法的制約の差異 (宮地, 1985) に着目して、両者の曖昧性解消に適用した。その文法的制約の大枠は次の通りである。

1. 連体修飾の制約
   (a) 関係節による修飾の禁止
   (b) ノ格句による修飾の禁止
   (c) 連体詞による修飾の禁止
2. 提題・取り立て助詞の制約
3. ヴォイスの制約
   (a) 受動態の禁止
   (b) 使役態の禁止
4. モダリティの制約
   (a) 否定表現の禁止
   (b) 意志性モダリティ表現の禁止
5. 文節分離の制約

連体修飾の制約に関して、(2) で見たように、慣用句「骨を折る」は連体修飾を受け付けないが、リテラル句なら問題ない。

ここで注意すべきは、橋本らの曖昧性解消法は慣用句にのみ適用される文法的制約を利用しているため、その制約に違反している句は高い精度でリテラル句として排除できる反面、制約違反の無いリテラル句は全く排除できない、という欠点である。明らかな制約違反の無いリテラル句には、奥 (1990) の報告にあるように、格フレーム制約の違反のチェックが必要である。

3 提案手法

橋本らによると、格フレームを有する述語慣用句はクラス B か C のみである。以下ではクラス B と C の格フレーム自動構築について述べる。

3.1 クラス B の格フレーム自動構築

クラス B は曖昧性が無いので、字面と構成素間の依存関係だけを手掛かりに正確に用例を収集できる。従って、収集した用例をそのまま河原らの格フレーム構築モジュールに入力すれば、§2.1 で述べた手続きに従って自動で慣用句格フレームが構築できる。次項で述べる曖昧性のある慣用句と違い、用例を慣用句とリテラル句に分ける必要は無い。

3.2 クラス C の格フレーム自動構築

§2.1 で述べたように、慣用句と直前格のペアによるクラスタリングを通して用法 (慣用句あるいはリテラル句) の違いを表すより、あらかじめ用例を慣用句とリテラル句に分けておいた方が良い結果が期待できる。

そこで本研究では、ある慣用句の格フレームを構築する際、もとになる用例集にはその慣用句の用例だけを含め、字面が対応するリテラル句の用例をあらかじめ区別しておく。つまり、用例収集の段階で曖昧性を解消しておく。この前処理により、クラス B と同じ格フレーム自動構築法を適用できる。

本研究では、用例の曖昧性解消に橋本らの手法を適用する。しかし、§2.2 の最後に述べた通り、この手法で高精度で曖昧性解消できるのは、明らかな文法的制約違反のあるリテラル句のみである。

そこで我々は、次のようなブートストラップ的な格フレーム自動構築法を考えている (図 1 を参照[2])。

[2]図 1 において、CF は格フレームを表す。また、実線矢印と Phase 番号が処理の流れを表し、点線矢印がどの手掛かり (文法的制約あるいは格フレーム) がどの処理で使われるかを表す。

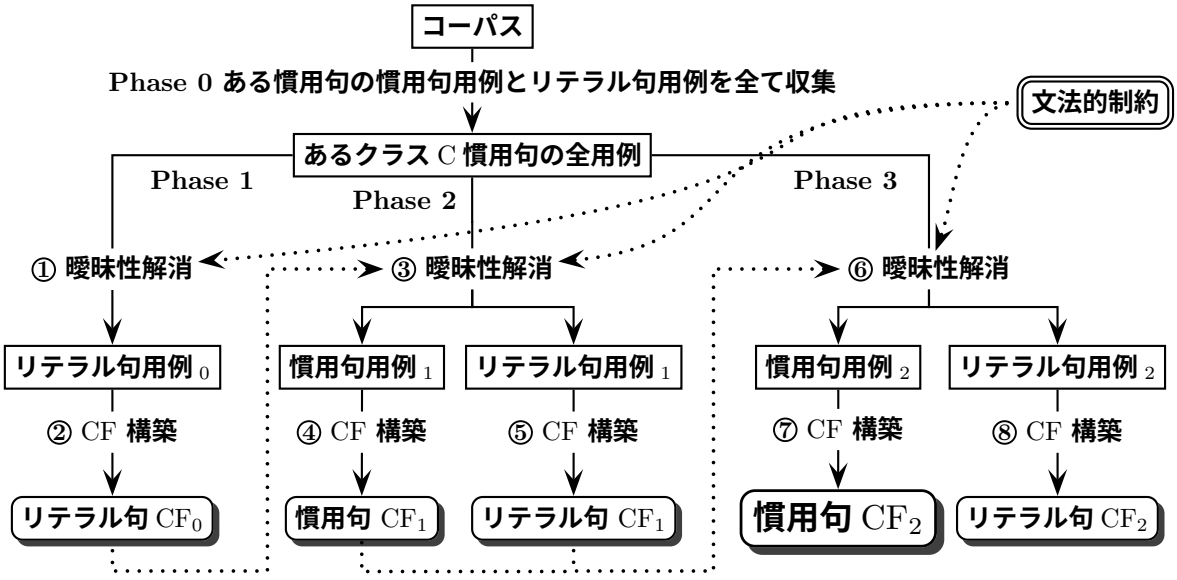

図 1: クラス C 慣用句の格フレーム構築手順

Phase 0) コーパスから、あるクラス C 慣用句 (例えば「骨を折る」) と字面が同じ句を含む文を全て収集する。つまり、この段階では、慣用句とリテラル句の区別を付けず、その句の用例を全て集める。

Phase 1) ① 次に、Phase 0 で集めた用例から、橋本らの文法的制約のみを手掛かりに、明らかな文法的制約違反のあるリテラル句用例のみをフィルタリングする。② 集めたリテラル句用例から、リテラル句の「格フレーム」を作る (リテラル句格フレームと呼ぶ)。

Phase 2) ③ そして、橋本らの文法的制約とリテラル句格フレームの 2 つを手掛かりに、全用例を再び慣用句用例とリテラル句用例に分ける。④⑤ 慣用句用例とリテラル句用例のそれぞれから、慣用句格フレームとリテラル句格フレームを作る。

Phase 3) ⑥ 橋本らの文法的制約とリテラル句格フレーム、そして慣用句格フレームの 3 つを手掛かりに、全用例をもう一度慣用句用例とリテラル句用例に分ける。⑦ 最後に、集めた慣用句用例から最終的な慣用句格フレームを作る。[3]

Phase 2 の③で全用例を慣用句用例とリテラル句用例に分ける際、文法的制約と Phase 1 のリテラル句格フレームを併用する。具体的には、まず文法的制約違反の有無をチェックし、違反があればリテラル句用例とする。違反がない用例は、その述語項構造が抽出され、リテラル句格フレームとの類似度がチェックされる。類似度が十分高ければリテラル句用例と判断され、そうでなければ慣用句用例と判断される。

Phase 3 の⑥では、文法的制約、リテラル句格フレーム、慣用句格フレームの 3 つが手掛かりとして使われる。具体的には、まず文法的制約違反の有無をチェックし、違反があればリテラル句用例とする。違反がない用例は、その述語項構造が抽出され、リテラル句格フレームと慣用句格フレームのそれぞれとの類似度がチェックされる。そして、より類似度が高い方の用例として判断される。

このブートストラップ的手法は、限られた手掛かり (文法的制約のみ) からスタートし、徐々に信頼できる手掛かり (リテラル句格フレーム、次いで慣用句格フレーム) を集めて、最終的に、高い精度で慣用句格フレームを構築することを狙っている。

4 対象とする慣用句とコーパス

本研究では、佐藤 (2007) に収録されている慣用句 (3,629 句) のうち、クラス B と C に属するものを対象として格フレームを構築する予定である。

佐藤 (2007) では A、B、C の 3 分類はなされていないので、あらかじめ分類しておく必要がある。ざっと分類してみたところ、A が 1,049 句、B が 2,035 句、C が 470 句、判断を保留したものが 75 句あった。

用例の収集元としては、Kawahara and Kurohashi (2006) に倣い、5 億文の Web コーパスを用いること

[3] **⑧で最終的なリテラル句格フレームも構築する。これは、我々の目的は慣用句格フレームの構築だが、後述するように、慣用句検出タスクによる評価実験の際、リテラル句格フレームも使用するためである。**

を考えている。

5 慣用句検出タスクによる評価

構築した慣用句格フレームは、橋本らと同様の慣用句検出タスクに適用して評価する予定である。このタスクは、入力された一文に対して、慣用句の有無を調べ、もしあれば、どの文節にどの慣用句 (の構成素) が含まれているかを出力するものである。

検出手法として表 1 の 8 つを用意し、その違いを見ることで慣用句格フレームの出来具合を評価する。例

表 1: 慣用句検出 8 つの方法

	慣用句 CF	リテラル句 CF	文法的制約
手法 0			
手法 1			✓
手法 2		✓	
手法 3		✓	✓
手法 4	✓		
手法 5	✓		✓
手法 6	✓	✓	
手法 7	✓	✓	✓

えば手法 0 では、何の手掛かりも用いず、全ての用例に対して「慣用句有り」と判断させる。一方手法 7 では、全ての手掛かりを用いて検出を試みる。

6 おわりに

本稿では、我々が計画している慣用句格フレームの自動構築について述べた。

§4 で述べた慣用句のうち、約 60 句のクラス B 慣用句について格フレームを構築した。クラス B の慣用句用例は正確に収集できるので、構築結果は非常に良好である。

一方、クラス C の格フレーム構築では、Phase 1 ①の曖昧性解消の善し悪しが大きな鍵を握っている。Web データからサンプリングした慣用句用例に対して、橋本らの文法的制約を用いてざっと曖昧性解消を試みたところ、橋本らの評価実験報告を若干下まわる結果が出た。これは、橋本らが新聞文を対象としていたのに対し、我々が Web 文を対象としているせいだと考えられる。つまり、Web 文は新聞文より口語的、あるいはくだけた文体なので、形態素解析／構文解析の段階で失敗することが多くなる。また、慣用句の文法的制約も Web 文の方がずっと緩い。本格的な格フレーム構築に入る前に、形態素解析器／構文解析器のロバスト化と文法的制約の Web 文に対しての最適化が必要である。

謝辞

慣用句データを提供してくださった名古屋大学大学院工学研究科電子情報システム専攻 佐藤理史研究室の方々に感謝申し上げます。

参考文献

Hashimoto, C., Sato, S., & Utsuro, T. (2006). Japanese Idiom Recognition: Drawing a Line between Literal and Idiomatic Meanings. In *COLING/ACL 2006 Poster*, pp. 353–360 Sydney.

Kawahara, D. & Kurohashi, S. (2006). Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of The 5th International Conference on Language Resources and Evaluation (LREC-06)*, pp. 1344–1347.

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.

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.

佐藤理史（編） (2007). 『基本慣用句五種対照表』. 名古屋大学大学院工学研究科 電子情報システム専攻 佐藤理史研究室.