慣用句の検出と格解析のための言語資源の構築*

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

我々は、日本語慣用句の検出・格解析のための言語資源を構築している。その言語資源は、辞書、検出・格解析器、用例集から構成され、フリーウェアとして公開する予定である。

慣用句の検出・格解析技術は正確な言語理解に欠かせない。例えば Excite の翻訳サイトで「その話が彼女の胸を強く打った。」を英語に翻訳させてみると、“The story strongly hit her chest.” と誤訳される。これは、「胸を打つ」を「感動させる」という意味の慣用句として検出し、「彼女」を経験者格、「その話」を対象格として解析できなかったためである。本研究の慣用句言語資源を機械翻訳器に組み込むことにより、上記の文を “She was very impressed by the story.” と翻訳することが可能となる。

以下では、本研究の中間報告を行う。

2　慣用句言語資源の全体像

2.1　対象とする慣用句

[6] の約 3,600 句の中から基本的な慣用句約 900 句を選定し、それらを対象とする。[6] では、基本慣用句リスト作成を目的として、小学生用辞典 2 つ、慣用句辞典 2 つ、慣用句研究文献 1 つの計 5 文献から慣用句を集めている。本研究の約 900 句は、3 つ以上の文献に記載されている句であり、最も重要なものといえる。

この約 900 句には文法的に互いに派生関係にあるもの（例えば「足元を見る」と「足元を見られる」）も含まれている。これらは、その関係を捉えた上で言語資源に組み込むべきだが、現段階ではそこまでは立ち入らず、それぞれを別項目として扱うことにした。

なお、[6] ではいわゆる故事成語とことわざも含まれているが、本研究ではこれらを区別せず、全て慣用句として扱う。約 900 句の中には故事成語が 10 句、ことわざが 38 句含まれている。

2.2　慣用句辞書

辞書には、検出・格解析器が必要とする情報、つまり、検出規則と格フレームが慣用句別に記載される。本研究の検出は、字面だけでなく、慣用句の意味と文字通りの意味を区別して行う。格解析は、ガ、ヲ、二等の格パターンだけでなく、格インスタンスの当該慣用句に対する妥当性も考慮する。

ここで重要なのは、慣用句によって必要となる辞書情報の量も質も異なるということである。[1] では、**形態変化の有無**と**曖昧性の有無**という 2 つの視点に基づいて慣用句を 3 つに分類した（図 1）。形態変化としては、述語部分の屈折、提題・取り立て助詞への変化、文節の分離などが含まれる。また、本研究における曖昧な慣用句とは、慣用句の意味と文字通りの意味の両方を持つ句のことである。

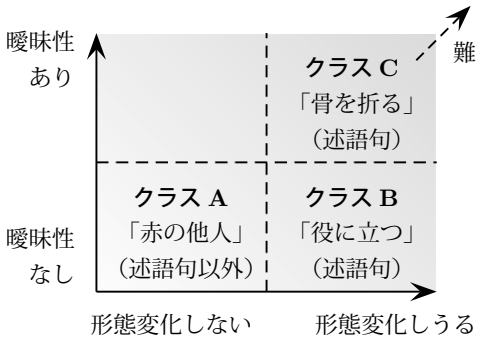

図 1: 慣用句の分類

本研究でも [1] の分類体系に従って慣用句辞書を構築する。この分類によって扱いが難しいクラスを特定することができ、結果、言語資源の構築と保守の手間を軽減できる。以下で述べる検出のための辞書情報の大枠も [1] に従う。

2.2.1　クラス A の辞書情報

クラス A の句は「赤の他人」のように形態変化も曖昧性もない。またほとんど全てが名詞句や副詞句、連体詞句などの非述語句であり、格解析の必要はない。

*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。

検出のための情報もその表記だけで済む。

2.2.2 クラス B の辞書情報

クラス B の句は曖昧性はないが、例えば「役に立つ」が「役には全く立たなかった」になるように、形態は変化しうる。検出の際はその変化に対応するため、慣用句を構成する内容語と変化しない機能語の依存関係に着目する。例えば「役には全く立たなかった」なら、「役」、「に」、「立た」の間に依存関係が存在するので、それらを慣用句「役に立つ」として検出する。よって、検出のための情報として構成語間の依存関係を与えておく必要がある。

クラス B は述語句がほとんどなので格解析の対象となる。本研究の格フレームは、[4] と同様の、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「役に立つ」の場合、(1) のようになる。

(1) [寄付 接待 贈与 ...] **ガ** [社会 営業 節税 ...] **ノ　役に立つ**

格フレームは [4] に倣い、次のようにして構築する。

(2) ① 当該慣用句の用例を収集する
② 各用例を構文解析し、述語項構造を抽出する
③ **ガ**、**ヲ**、**ニ**等の格マーカー毎に格インスタンスをクラスタリングする

2.2.3 クラス C の辞書情報

「骨を折る」等のクラス C 慣用句は、形態変化するだけでなく曖昧性もあり、最も扱いが難しい。検出のための辞書情報としては、クラス B と同様、構成語間の依存関係を与えておく必要がある。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の手掛かりも必要となる。

[1] では、「骨-を-折る」のような「名詞-助詞-動詞」型の慣用句を対象に、慣用句にのみ当てはまる文法的制約 (3) を手掛かりに曖昧性解消を試みた。

(3) a. 名詞構成素への連体修飾の制約
b. 提題・取り立て助詞の制約
c. ヴォイス（受け身、使役）の制約
d. モダリティ（命令、許可、禁止、意志、否定形など）の制約
e. 文節分離の制約

これは、慣用句と文字通りの意味の句（リテラル句と呼ぶ）とでは許される文法的操作の範囲に差があるという仮説に基づいている。例えば「骨を折る」が慣用句として使われる場合、連体修飾は受けつけない。(4b) は連体修飾の例だが、(4a) と違い、文字通りの意味にしか解釈できない。

(4) 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。

(3) の制約全てが全ての慣用句に当てはまるわけではない。例えば「骨を折る」は、「骨を折らせる」のように使役形になっても慣用句の意味を持ちうる。よってクラス C の辞書エントリには、慣用句毎に当てはまる制約それぞれが個別的に列挙される。クラス C の検出（曖昧性解消）については §4 でさらに議論する。

クラス C 慣用句もクラス B と同様、述語句がほとんどであり、格解析の対象となる。格フレームはクラス B と同様のものだが、構築の際、より手間がかかる。具体的には、(2)–① の用例収集の際、文字通りの意味で用いられている用例は除外し、慣用句の用例だけを選別する、つまり曖昧性解消する必要がある。これについては §4 でさらに議論する。

2.3 慣用句検出・格解析器

上述の辞書情報は構文解析器 KNP [5] に組み込まれる。つまり本研究では KNP を慣用句検出・格解析器として用いる。

検出用の辞書情報は KNP の規則（形態素列・係り受けパターン）として記述される。(3) の曖昧性解消情報の場合は、既に KNP に組み込まれている、連体修飾、ヴォイス、モダリティ表現等の認識機能も利用して制約を記述する。

格フレームは、既に KNP にある [4] のものと同様の形式にして、既存のものに付け足す形で実装する。

慣用句検出の結果は、当該慣用句の構成要素を含む文節に、各慣用句に与えられている ID を付与する形で出力される。格解析の結果は KNP の格解析結果出力形式に準拠する。

2.4 慣用句用例集

慣用句用例集は、慣用句研究の基礎資料として広く利用されることを期待して収集、公開する。

用例収集の対象とするのは、約 900 句全てが望ましいが、クラス C の句のみに限定する。用例数として、1 句あたり、慣用句用例（正例）と文字通りの意味の用例（負例）を各々50 文ずつ収集する予定である。また、構文解析等のアノテーションはしない。

3 進捗状況

慣用句言語資源は次の段階を経て公開に至る。

(5) 1 慣用句の A、B、C 分類 ……終了
2 検出規則の開発、実装
3 格フレームの構築、実装
4 用例の収集、整備

現在、慣用句の分類が終了している。分類は著者 1 名と作業者 1 名が並行して行った。食い違うものについては話し合いにより、標準的な日本語話者の直観により近い方を選んだ。分類の結果、クラス A が 20.0%、B が 63.4%、C が 16.6%という内訳になった。

残る作業のうち、クラス A と B に関わるものは全て容易に行える。一方クラス C に関しては、検出規則の開発、格フレームの構築、用例の収集全てが野心的な課題といえる。これについて §4 でさらに議論する。

4 課題

ボトルネックとなっているのはクラス C 慣用句の曖昧性である。これがうまく解消できれば、クラス C の格フレーム構築も用例の収集も容易に行える。

語義曖昧性解消法の主流は大量のデータから学習した分類器を用いるものだが、クラス C 慣用句の全てに十分な学習データを用意するのは大変手間がかかる。そこで [1] では、(3) にある少数の文法的制約のみを慣用句に与え、それにより曖昧性解消を試みた。[1] のシステムは制約違反のある用例を負例として排除する形で曖昧性解消する。クラス C 慣用句 34 句を対象に新聞文を用いて小規模な実験をしたところ、排除した例は全て負例だが全負例 42 例のうち 15 例しか排除できないという高精度、低カバー率の結果に終わった。

[1] では、曖昧性解消に失敗した 27 の負例のうち 5 例には格フレームが有効であると報告している。しかし §2.2.3 で述べたように、クラス C の格フレームを構築するには用例の曖昧性解消が必要となる。つまり、曖昧性解消と格フレーム構築は相互依存関係にある。

そこで、高精度低カバー率の (3) の制約を用いたブートストラップ的な格フレーム構築法を考えている（図 2）。

Phase 0) あるクラス C 慣用句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を区別しないで用例を全て集める。

Phase 1) ❶ Phase 0 の全用例から (3) の制約のみを手掛かりに明らかな制約違反のあるリテラル句用例をフィルタリングする。❷ 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 CF と呼ぶ）。

Phase 2) ❸ 制約とリテラル句 CF 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。❹❺ それぞれの用例から慣用句 CF とリテラル句 CF を作る。

Phase 3) ❻ 制約、リテラル句 CF、慣用句 CF の 3 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。❼ 慣用句用例から最終的な慣用句 CF を作る。

❸で用例を分ける際、制約と Phase 1 のリテラル句 CF を併用する。具体的には、まず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 CF 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。❻では、制約、リテラル句 CF、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 CF と慣用句 CF 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（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 CF、次いで慣用句 CF）を集めて、最終的に高い精度で慣用句 CF を構築することを狙っている。

この手法の成否は❶の結果の善し悪しにかかっている。Web から取ってきた慣用句用例に対して (3) の制約で曖昧性解消を試みたところ、[1] の報告を若干下まわる結果が出た。これは [1] が新聞文を対象としていたのに対し、我々が Web 文を対象としたせいだと考えられ。つまり、Web 文は新聞文よりくだけた文体が多いので構文解析の段階で失敗しやすい。また、文法的制約も Web 文の方がずっと緩い。Web 文から格フレームを構築する場合は、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。

結局本研究では、(3) の制約と格フレームを併用して曖昧性解消を行う。しかし、これで全て句の曖昧性解消が可能というわけではない。例えば慣用句「汗を流す」は、制約、格フレームともに、慣用句とリテラル句とで大きな違いはなく、曖昧性解消の役に立たない。これらに対しては分類器を学習しておく必要がある。

重要なのは、クラス C 慣用句を一括りにするのではなく、制約と格フレームで対応できるものと、さら

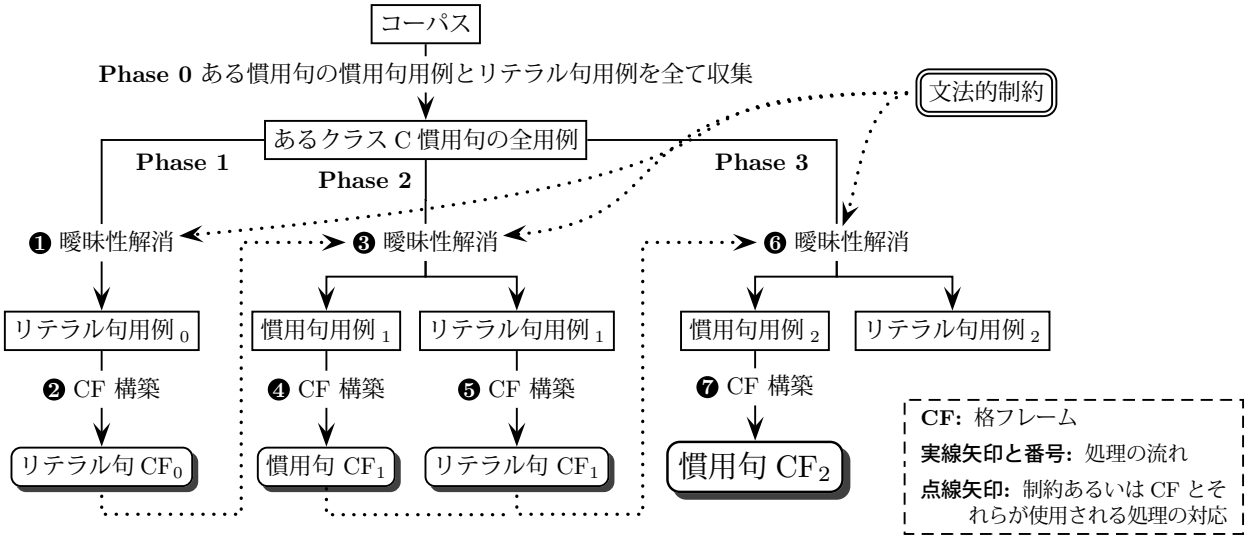

図 2: クラス C 慣用句の格フレーム構築手順

に分類器の学習が必要になるものとを分けることである。これにより言語資源の構築と保守の手間をさらに軽減できる。

上記により検出のための制約と格フレームが完成すれば、その副産物として得られた慣用句用例を用例集として使うことも、制約と格フレームが組み込まれた検出器によりさらに用例を収集することも容易である。用例集は最終的には人手でチェックして仕上げる。

5 関連研究

日本語慣用句言語資源の研究は [3] や [2] など以前にもあった。しかしフリーウェアとして公開されているわけではない。

また慣用句言語資源構築の方法論に関しても、従来のものは、構築と保守の手間を軽減する、形態変化の有無と曖昧性の有無という分類のための有用な視点が欠けている。

6 おわりに

本稿では我々が現在構築している慣用句言語資源について中間報告を行った。その言語資源はフリーウェアとして公開する予定であり、完成すれば、日本語の基本的な慣用句の検出と格解析が可能となる。検出は、字面だけでなく、慣用句の意味と文字通りの意味を区別して行う。格解析は、ガ、ヲ、二等の格パターンだけでなく、格インスタンスの当該慣用句に対する妥当性も考慮する。本研究の特長は、形態変化の有無と曖昧性の有無という 2 つの軸に基づいて慣用句を分類する点にある。クラスによって必要な情報も構築の手間も異なるため、全慣用句の中から難しいクラスを特定することで、構築と保守の手間を軽減できる。

現在、[6] の慣用句リストを分類するところまで終了した。残る作業で難しいのはクラス C に関するものだけである。その展望については §4 で議論した。

本研究の慣用句言語資源は 2009 年度末に公開する予定である。

謝辞

慣用句リスト [6] を提供してくださった名古屋大学大学院 佐藤理史先生に感謝申し上げます。

参考文献

[1] Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. Detecting japanese idioms with a linguistically rich dictionary. *Language Resources and Evaluation*, Vol. 40, No. 3–4, pp. 243–252, 2006.

[2] Kosho Shudo, Toshifumi Tanabe, Masahito Takahashi, and Kenji Yoshimura. MWEs as Non-propositional Content Indicators. In *the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 32–39, 2004.

[3] 奥雅博. 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, Vol. 31, No. 12, pp. 1727–1734, 1990.

[4] 河原大輔, 黒橋禎夫. 格フレーム辞書の漸次的自動構築. 自然言語処理, Vol. 12, No. 2, pp. 109–131, 2005.

[5] 黒橋禎夫, 長尾真. 長い日本語文における並列構造の推定. 情報処理学会論文誌, Vol. 33, No. 8, pp. 1022–1031, 1992.

[6] 佐藤理史（編）. 基本慣用句五種対照表. 名古屋大学大学院工学研究科 佐藤理史研究室, 1 2007.