様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 6 月 3 日現在

研究種目：若手研究(B)
研究期間：2007～2009
課題番号：19700141
研究課題名（和文）　日本語慣用句の検出と格解析のための言語資源の構築
研究課題名（英文）　Construction of language resources for the detection and case analysis of Japanese idioms

研究代表者 橋本 力（HASHIMOTO CHIKARA）
独立行政法人　情報通信研究機構・知識創成コミュニケーション研究センター言語基盤グループ・専攻研究員
研究者番号：00402800

研究成果の概要（和文）：本研究では、コンピュータによる言語理解の実現に向けて、「骨を折る」などの慣用句を文章中から自動検出するための研究を行った。検出の際は文字通りの意味と慣用的意味を区別する。例えば「骨を折る」なら、「骨折する」と「苦労する」の 2 つの意味を区別する。本研究では、コンピュータが慣用句検出を自動学習するのためのデータを構築し、過去に提案されてきた単語の意味的曖昧性技術に基づく、独自の慣用句検出手法を開発した。

研究成果の概要（英文）：This study addressed the automatic detection of Japanese idioms with their semantic ambiguity resolved. We constructed a data by which a computer system learned the idiom detection, and developed the automatic idiom detection method exploiting the data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,300,000	0	1,300,000
2008 年度	1,300,000	390,000	1,690,000
2009 年度	700,000	210,000	910,000
年度			
年度			
総　計	3,300,000	600,000	3,900,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：自然言語処理, 日本語, 慣用句, 慣用表現, 検出, 曖昧性解消, コーパス

１．研究開始当初の背景

コンピュータによる言語理解の実現のためには数多くの言語処理技術が構成要素として必要であるが、そのうちの一つとして慣用句検出が挙げられる。慣用句検出とは、文章中に存在する慣用句を自動検出する課題であるが、この課題の難しさは、慣用句に相当する文字列（例えば「骨を折る」）が慣用句としての意味（「苦労する」）だけでなく、文字通りの意味（「骨折する」）も表しうる点にある。つまり慣用句検出には、意味的曖昧性の解消が必要となる。

しかし、これまでの慣用句研究のほとんどは慣用句として使われる文字列の収集に留まっていた。英語を対象とする慣用句検出に関する研究はわずかに存在するが、日本語の

慣用句検出研究は存在しない。本研究はこの隙間を埋めるものとして意義がある。

本研究が採用するアプローチは、大規模データに基づく統計的手法によるものである。具体的には、人手で正解が付与された大規模なデータを教師信号とし、入力に対して望むべき出力を得るようなシステムを自動学習するという枠組みである。形態素解析、構文解析、意味解析、照応解析など、多くの自然言語処理技術がこの枠組みによって大きく進展した。本研究が取り組む課題、慣用句検出においてもこの枠組みが奏功する可能性が高い。本課題で必要となるデータは、文字通りの意味か慣用句かを示すラベルが付与された、慣用句候補の文字列を含む文の集合（コーパス）である。

しかし、過去このようなコーパスを構築した研究は数少ない。英語では、次に挙げる研究がある。

（1）Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recoginition of nonliteral language. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pages 329–336.

（２） Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions, pages 41–48.

日本語では尾島、佐藤が小規模なコーパスを構築した。

（3）尾嶋, 佐藤, 宇津呂："日本語慣用句用例データベースの構築法", 言語処理学会第 12 回年次大会, pp. 456 - 459. (2006).

本研究で構築するコーパスは１０万文規模のものであり、世界最大規模と言える。慣用句トークン検出手法に関しても、過去提案されたものはわずかである。英語では次の研究がある。

（4）Julia Birke and Anoop Sarkar. 2006. A clustering approach for the nearly unsupervised recoginition of nonliteral language. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), pages 329–336.

（５） Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In Proceedings of the Workshop, COLING/ACL 2006, MultiwordExpressions: Identifying and Exploiting Underlying Properties, pages 12–19, July.

（６） Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions, pages 41–48.

日本語では次の研究がある。

（７） Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. 2006. Japanese Idiom Recognition: Drawing a Line between Literal and Idiomatic Meanings. In The Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006) Poster, pages 353–360, Sydney, July.

本研究では、従来の単語の意味的曖昧性解消手法に依拠しつつ、より精度を上げるための新たな提案をする。

２．研究の目的

本研究では、よく使われる基本的な慣用句 146 句を対象に、10 万文規模の慣用句コーパスを構築する。対象とする 146 句は次のようにして選ばれた。

（1）基本慣用句五種対照表の約 3,600 句の中からより基本的な句を抽出する。基本慣用句五種対照表では、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。我々はより基本的な句として、基本慣用句五種対照表中の 5 文献中、3 つ以上の文献に記載されているもの 926 句を抽出した。

（2）（1）の結果から、曖昧性があると認められる句を、複数人による判断結果を踏まえ、研究代表者が選定する。結局、146 句が曖昧性ありと認められた。

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

本研究の慣用句コーパスは慣用句の正確な検出技術、とりわけ、正確な曖昧性解消技術の開発のために構築されている。そのためコーパス中の各用例には、用例中の慣用句相当の文字列が慣用句として用いられているのか、あるいは文字通りの意味として用いられているのかを表すラベルが付与されてい

る。以後、前者のタイプの用例を正例、後者のタイプを負例と呼ぶ。具体的には、一行が一用例の情報に対応し、各行は下記の 4 つの情報から成る。

① ラベル：正例なら「i」、負例なら「l」
② ID：各慣用句に与えられた 4 桁の数字。
③ 表記：対象とする慣用句の表記。
④ 用例：用例文そのもの

用例数は、原則として、慣用句ごとに、正例負例併せて 1,000 用例になるまで収集する。1,000 文に至らない場合はなるべく多く収集する。用例は全て Web コーパスから収集する。　下記に慣用句「ごまをする」の負例の例を挙げる.

l 1417 ごまをする すり鉢でごまをすり、ごま油を...

コーパスの構築に加えて、本研究では、慣用句検出技術を開発する。本研究では、上記の慣用句 146 句を対象とする。タスク設定は次の通りである。各慣用句について、上記で得た正例と負例を、正例／負例ラベルを取り除いた状態でシステムに入力し、システムは各用例に対して、正例か負例かを判断する。

３．研究の方法

下記の手順で慣用句コーパスを構築する。

（１）Web コーパスから慣用句相当の文字列を含む文を(正例負例関係なく)収集する。具体的には、慣用句の構成語が全て正しい係り受け関係で出現している文を自動で収集した。その際、係り受け解析器 KNP を使用した。

（２）収集された文を人手で正例と負例に分ける。この判断は上記の意味参照リストに基づいてなされる。正例負例併せて 1,000 文になるまでアノテーションするが、長い文から優先的にアノテーションした。また、慣用句相当文字列として不的確な誤って自動収集された用例と、前後の文脈無しには正例負例の判断がつかない用例はアノテーション対象から除外した。この作業は 2 名がおよそ 230 時間かけて完了した(460 人時)。

慣用句検出に関して、本研究では、次のような機械学習に基づく手法を採用した。学習法は SVM (2 次の多項式カーネル。TinySVM を使用) で、使用した素性は下記の通りである。

① f1:前後 3 語（機能語も含む）の品詞
② f2:前後 3 語(機能語も含む)の表出形(形態素解析器 JU-MAN でいう「見出し」)
③ f3:文中の全内容語の原形
④ f4a:係り元形態素（慣用句の先頭文節に係る文節の中で一番最後に現れる文節における、最初の形態素）の原形
⑤ f4b:係り元形態素の品詞
⑥ f5a:係り先形態素（慣用句の最後の文節が係る文節の中で一番最初に現れる文節における、最初の形態素）の原形
⑦ f5b:係り先形態素の品詞
⑧ f6:文中の全内容語の JUMAN カテゴリ
⑨ f7:文中の全内容語の JUMAN ドメイン

これらの素性の抽出処理には JUMAN と KNP を用いた。f4 と f5 は、の Syntactic Relations 素性を慣用句用に設計し直し、かつ、より単純にしたものである。例えば次の「胸を打つ」の用例の場合、f4 は「聴衆」の原形と品詞、f5 は「歌声」の原形と品詞になる。

（聴衆/の）（胸/を）（打つ）（美しい）（歌声）

f6 と f7 は JUMAN が出力する情報の一つで、前者が単語の上位概念に、後者が単語の属する分野あるいはトピックに概ね対応する。例えば「聴衆」の場合、JUMAN カテゴリは「人」、JUMAN ドメインは「文化・芸術あるいはメディア」である。「歌声」なら、JUMAN カテゴリは「抽象物」、JUMAN ドメインは「文化・芸術あるいはレクリエーション」である。

慣用句検出実験では、正例と負例がともに 50 用例以上利用可能な慣用句 93 タイプを対象として、慣用句タイプごとに SVM モデルの構築と評価実験を行う。評価指標は Accuracy を採用する。

Accuracy=正しく曖昧性解消できた用例数/全用例数

ベースラインは、慣用句コーパスにおいてより用例数が多い方に一律に解釈した場合とする。

Baseline Accuracy=max(正例数;負例数)/全用例数

これは慣用句ごとに決定される。例えば、慣用句コーパスにおいて、正例数が 600 で負例数が 400 なら、その慣用句の Baseline Accuracy は 60%となる。タイプごとの Accuracy と Baseline Accuracy は、その用例集の学習用／評価用への分割の仕方を 10 通りに変化させて得た 10 の値の平均として求める（10 分割交差検定のスタイル)。

実験対象の慣用句全体での Accuracy と Baseline Accuracy も求める。これは、慣用句タイプごとの Accuracy を合計し、慣用句のタイプ数（93）で割ったもの（macro-averaged）である。また、Relative Error Reduction(RER)を次の式に基づいて算出する。

RER=(ベースラインのエラー率-システムのエラー率)/ベースラインのエラー率

慣用句全体の RER は全体の Accuracy と Baseline Accuracy から上の式で直接算出する。

４．研究成果

本研究で構築した慣用句コーパスは、最終的に、用例の総数が 113,460 文となった。68 慣用句に対して 1,000 用例以上構築できたが、100 用例未満の慣用句も 17 ある。一用例

の平均単語数は 46 語と比較的長い。慣用句コーパスは長い文から優先的に 1,000 用例に至るまでアノテーションして構築しているので、他のコーパス（新聞記事コーパスや WWW コーパス）と比べて長い文が多い。正例負例分類作業の一致率（Kappa 統計量）を計るべく、全 113,460 用例から 1,421 用例をサンプリングして、グループ乙の 2 名に同様の作業を行わせた。結果、グループ乙の 2 名の間の一致率は 0.8519 と高く、判断の揺れが少ないことが分かった。　以下に「胸を打つ」と「ごまをする」の最も長い正例と負例、最も短い正例と負例をそれぞれ挙げる。「胸を打つ」の最も長い負例は自動文区切りの結果が間違っているが、手作業修正等は今のところしていない。

（1）i3193 胸を打つこのアラビア語版『武士道』を讀んだアラブの大使や外交官から日本人の精神が良く分かったと言う感想を貰ったが、「切腹はわが国の中世にはじまって、武士がその罪をつぐない、過ちを謝し、恥をまぬがれ、友人につぐない、そして自分の誠実を証明する方法であった」という解説に、当時のエジプト大使館の情報参事官から「切腹など全く野蛮な行為と思っていたが、腹を切ったらさぞ痛からうに、それをあえて実行するサムライの精神の高さに胸を打たれた」という感想を貰ったことも懐かしい思い出である。

（2）l3193 胸を打つと言うのも、自転車で横断歩道を渡ろうとして、左折した車とぶつかってこけてしまったんです　その時に自転車のハンドルで胸を打ち、すぐ救急車を呼ばれて…　歩いて３分くらいの病院なのにサイレン鳴らされました（＾ー＾；　幸い大事には至ず、車に乗っていた方も誠意ある対応をして下さったんですが、その後警察で事情聴取や今後どうするかなどの説明を受け大変な一日でした！

（3）i3193 胸を打つよくぞここまでなキャラの立った登場人物が自分勝手に行動すればするほど笑いが起こり、かと思うと一点、熱い友情が展開されて胸を打つ。

（4）l3193 胸を打つ顔を天に向けようともせぜ、自分の胸を打ち叩き「こんな罪人の私を憐れんで下さい」と祈る姿をイエス様は義と認めて下さったのです。

（5）i1417 ごまをするただし、１５６２年にグレシャムは１５６２年に、同一の額面価値で流通する素材価値を異にする２種類の貨幣が存在すると劣悪な貨幣が流通に残り、優良な貨幣は駆逐されるという「グレシャムの法則」を発表していることから、女衒がしたたかであったように、ＩＴ興行師もメーカーにごまをすり、政府の省庁にこびを売り、知性が無くても生命力だけで、目新しい言葉のつまみ食いで、悪毒、図々しく、虚勢で生き続けることだろう。

（6）l1417 ごまをする煮た大豆をつぶすには、ミンチみたいな器具があればいいのですが、ない　場合は、ごまをするもので潰すか、もっと簡単な方法としては、ビニール　袋に大豆を入れ、封をしてタオルをかけてその上から瓶でこするようにす　るといいでしょう。

（7）i1417 ごまをする上にごまをする小役人タイプ。

（8）l1417 ごまをするごまをすり調味料とあえる

完成した慣用句コーパスは、http://openmwe.sourceforge.jp/にて、BSD ライクなライセンスのもと、無料配布している。

本研究で開発した手法を慣用句検出実験によって評価した結果について述べる。提案手法の全体の Accuracy は 89.19%、Relative Error Reduction は 59.07%、ベースラインは 73.60%となった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

[1] Chikara Hashimoto and Daisuke Kawahara. Compilation of an Idiom Example Database for Supervised Idiom Identification. Language Resources and Evaluation. Volume 43, Number 4, pp.355-384. 2009. 査読有

[2] Chikara Hashimoto, Satoshi Sato, and Takehito Utsuro. Detecting Japanese Idioms with a Linguistically Rich Dictionary. Language Resources and Evaluation: Special Issue on Asian Language Technology, Volume 40, Number 3-4, pp.243-252. 2006. 査読有

〔学会発表〕（計７件）

[1] Chikara Hashimoto. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD incorporating Idiom-Specific Features. EMNLP2008: Conference on Empirical Methods in Natural Language Processing, pp.991-1000. 2008/10/27. The Hilton Prince Kuhio Hotel, Waikiki, Honolulu, Hawaii.

[2] 橋本力. 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. 電子情報通信学会研究会 言語理解とコミュニケーション研究会, pp.1-6. 2008/7/17. 公立はこだて未来大学, 北海道函館市.

[3] 橋本力. 慣用句の検出と格解析のための言語資源の構築. 言語処理学会第１４回年次大会発表論文集, pp.1148-1151.

2008/3/20. 東京大学, 東京都目黒区.
[4] Chikara Hashimoto. Japanese Idiom Recognition: Drawing a Line between Literal and idiomatic Meanings. COLING/ACL 2006 Poster, pp.353-360. 2006/7/17. Sydney Convention and Exhibition Centre, Sydney, Australia.
[5] 橋本力. 自動検出のための慣用句の分類と語彙的情報. 第 173 回 自然言語処理研究会. 2006-NL-173, pp.59-66. 2006/5/19. 東京農工大学, 東京都小金井市.
[6] 橋本力. 依存構造照合に基づく慣用句自動検出. 言語処理学会第１２回年次大会発表論文集, pp.829-832. 2006/3/16. 慶応義塾大学,神奈川県横浜市.
[7] 橋本力. 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回年次大会発表論文集, pp.825-828. 2006/3/16. 慶応義塾大学,神奈川県横浜市.

〔その他〕
ホームページ等
http://openmwe.sourceforge.jp/

６．研究組織
(1)研究代表者 橋本 力(HASHIMOTO CHIKARA)
独立行政法人　情報通信研究機構・知識創成コミュニケーション研究センター言語基盤グループ・専攻研究員

研究者番号：00402800

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：