イディオムの異形規則を利用したイディオム検索システムの構築

金平昂† 平尾一樹† 竹内孔一† 影浦峡♯

†岡山大学大学院自然科学研究科
♯東京大学大学院教育学研究科

1 はじめに

本研究では，英文テキスト中に現れる多様な異形を含む英語イディオムを，辞書に登録された標準形と自動的にマッチさせる手法について報告する．開発にあたっては，イディオムの異形を整理し分析した結果 [3] を利用した．金平ら [3] によると，異形の大部分は語の「置換」と「挿入」によるものである．したがって，置換と挿入による異形をもつイディオムの検索をカバーすれば，実用上は十分なシステムを実現できる．本研究では，このうち挿入による異形を扱う．置換による異形の処理には高精度のシソーラスが必要となり，具体的に入手可能なシソーラスとその活用方式も含めて，次のステップで扱う予定である．

本研究の背景には，我々が進めている，オンラインのボランティア翻訳者を支援する統合的なレファレンス・ツール／環境の高度化がある [1]．翻訳者は，既存の高品質辞書に掲載されているイディオムにほぼ満足しているが，その検索機能には満足していない [1][2]．紙の辞書を用いる場合，様々なイディオムについて，核となる単語を推測して引かなければ検索することもできず，イディオムが見落とされがちである．この状況は，電子辞書やオンライン辞書でもほぼ同様であり，キーとなる語を AND 検索しなくてはイディオムはわからない．

さらに，機械翻訳システムでも，イディオムの処理は不十分である．例えば，

(1) He said that with his tongue in his cheek.

(2) He said that with his big fat tongue in his big fat cheek.

では，“with one’s tongue in one’s cheek”（意味：あざけって）というイディオムが存在する．(1) を正しく処理できるシステムはあるが，(2) を正しく処理できるシステムは，我々が調べた範囲では存在しなかった [4]．

一方，イディオム検索が実際に必要とされる場面では，ある範囲にイディオムが存在すると予測されながら，正確にどこかわからないということが頻発する．そのため，多くの翻訳者が AND 検索は不十分であると感じており，電子テキストに対して，翻訳者がイディオムがありそうだと推測した領域をテキスト中で指定すれば，該当するイディオム候補を提示するシステムを望んでいる．

本研究では，既存のリソースや手法の利用に関して，POS-tagger や形態素解析器は使うが，構文解析器は使わないことにした．構文解析器は，実用的なシステムに用いるには，まだ精度とロバストさが十分でないと判断したためである．したがって，品詞情報と有限オートマトン以上の計算クラスは想定しないかたちで手法を定義し，システムを構築することとした．

この基準に従い，本研究では，具体的なリソースとして，品詞や活用形を扱うため，Tree-Tagger[6] を利用した．したがって，本システムの精度は，出発点において Tree-Tagger に依存してしまうことになる．なお，イディオムの基本データとしては，グランドコンサイス英和辞典 [7] を用いた．

2 イディオム検索

2.1 概要

本研究で作成するシステムは，ユーザが入力として英文を入力すると，正解候補イディオム（文中で使われているイディオム，または翻訳の決定に役立つイディオム）を検索し，表示する．翻訳者が翻訳を確定する際には，最終的には正解でなくても，誤訳の可能性を潰すために，翻訳者にとってあり得そうな範囲のイディオムを検討する．ここでは，これを翻訳の決定に役立つイディオムと呼ぶ．この範囲は，実際には個別の翻訳者によって異なるが，我々のシステムを評価する際には，翻訳の決定に役立つイディオムを近似的に定義する．これついては，３節で述べる．

検索の流れとしては，まずイディオムを辞書データとマッチさせるために，入力文に対して語の標準化を行ない，入力文と辞書データの語を整合させる．次に標準化した後の文に対して辞書データと表層のマッチングを行ない，可能性のあるイディオムを過剰に検索し，イディオム候補とする．そして，イディオム候補に

対して品詞情報を用いたフィルタリングを行ない，正解候補イディオムを出力として表示する．図 1 に検索の流れを示す．

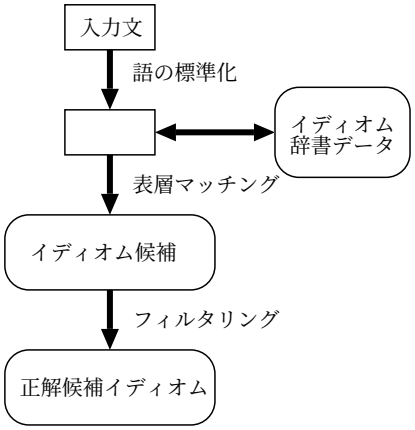


図 1: 検索の流れ

2.2 語の標準化

テキスト中と辞書中では語形が異なる場合が多くある．例えば，テキスト中では，“took his seat” と使われていても，辞書中では，“take one’s seat” として登録されている．したがって，テキスト中の語を辞書中の語形に揃える処理を行なった．語形が異なる場合は，(1) 動詞の活用形の違いや代動詞 (do，doing) での記述，(2) 名詞の複数形／単数形の違い，(3) 冠詞の違い，(4) 代名詞の表記の違い（his，myself が one’s，oneself で表記されている），がある．その他にハイフン (-) で連結している語はハイフンで区切り，別々の単語として扱うことにした．表 1 に語の標準化について示す．表 1 中の “表層” とは，入力語そのものの形のことである．辞書データでは “動詞” について，(a)“基本形” の場合，(b)“be able to do” のような “do” で表されている場合，(c)“cannot help doing” のようなで “doing” で表されている場合がある．“冠詞” については，“come to a point”（意味：獲物の在処を示す）と “come to the point”（意味：要点に触れる）のように，“a” と “the” で意味が異なるイディオムは検索対象に定めている．“my” や “myself” 等は辞書ではそれぞれ，“one’s”，“oneself” の形で記述されていることがある．

2.3 表層マッチング

表層マッチングでは，テキスト中に存在する可能性のあるイディオムを過剰に検索し，イディオム候補とする．辞書データに登録されているイディオムで，各

表 1: 語の標準化

入力語	パターン
(1) 動詞	表層，基本形，do，doing
(2) 複数形	表層，単数形
(3) 冠詞	a, an, the
(4)my, his 等	表層，one’s
(4)myself, himself 等	表層，oneself

イディオム毎にイディオムを構成する単語がすべてその順番で入力文の単語とマッチした場合にイディオム候補とする．辞書データには，“make A of B” や “have … in” ように，“A，B” や “…” を含むイディオムが存在する．ここでは，“A，B” や “…” はワイルドカード（何がマッチしてもよい）として扱う．図 2 にマッチング例を示す．この例では，“have one’s eye on” はすべての構成単語がマッチするのでイディオム候補である．一方 “have one’s ears on” は “ears” がマッチしないのでイディオム候補ではない．

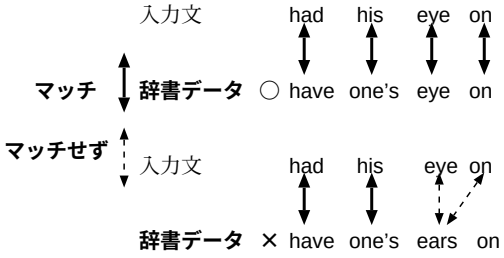


図 2: 表層マッチング

2.4 品詞情報を用いたフィルタリング

次にイディオム候補から正解候補イディオムを抽出するためのフィルタリングを行なう．利用情報としては，語の品詞情報を利用する．フィルタリングには，挿入語の品詞とイディオムの挿入箇所の前後の品詞パターンを解析した結果 [3] に我々がデータから考えた規則を加えたものを利用した．図 3 にフィルタリングの例を示す．入力文：“I make a habit of stretching after I wake up.” からは，表層マッチングにより，“make a habit of doing”，“wake up”，“make after”，がイディオム候補として検索される．ここで，金平ら [3] の解析結果より，“動詞” と “前置詞” の間に挿入される語は “副詞” と “形容詞” のみ認められている．ここで入力文中の，“make” と “after” の間の “a habit of stretching” は，この条件を満たさない．従って，“make after” は正解候補イディオムではないとみなされる．

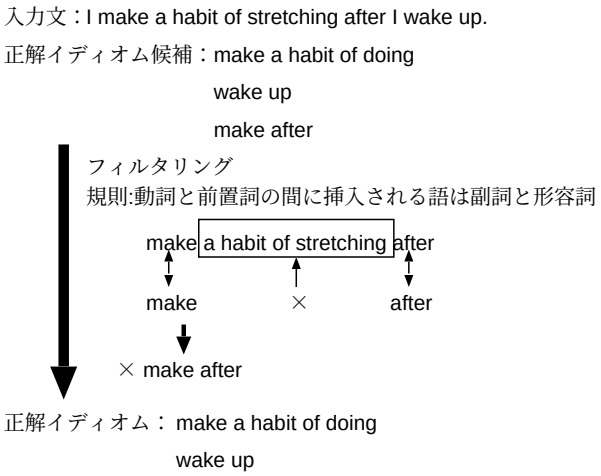


図 3: 品詞情報を用いたフィルタリング

2.5 インターフェース

実際にシステムのインターフェイスを図 4 に示す．ユーザは英文を入力し，ウィンドウ幅を決める．ウィンドウ幅とは，検索範囲のことであり語数で決定する．次に検索開始を選択するとイディオムとその意味が表示される．このとき，マウスポインタを表示されているイディオムに当てると，入力文中のイディオムである語が反転される．図 4 では，入力 “I decided to take the wild plunge and buy the car I had my eye on.” から，“take the plunge”，“have one’s eye on” がイディオムとして検索される．

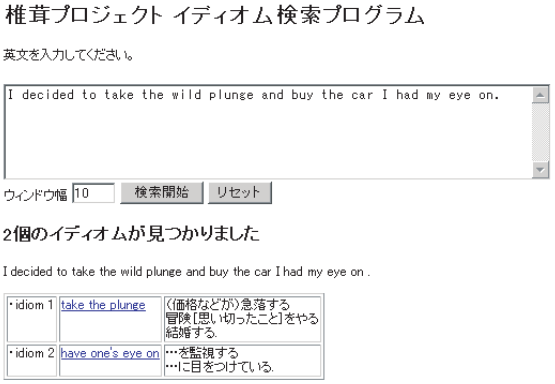


図 4: イディオム検索システム

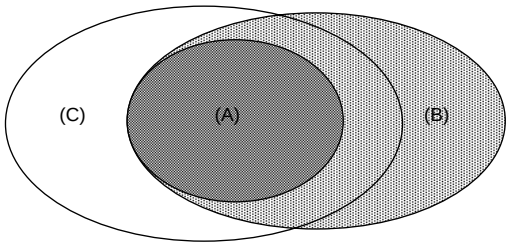
3 評価実験

実験の目的は，ここでは (1) 入力文の語のパターン化によって，どの程度辞書データのイディオムを検索可能であるかを明らかにする，(2) 品詞情報を用いることの有効性を確認する，(3) 実際の記事に対する有効性を確認する，について評価実験を行なう．

- 実験方法
  実験は対象データの各文に対して正解候補イディオムがどのくらい検索可能であるかを調べた．実験の対象データとしては，(a) 金平ら [3] で用いられている英語母語者によって対訳辞典 [5] から作成されたイディオムの異形を含む文（100 文），(b) 実際の英語記事 (20 記事)，を用いた．(b) の実際の英語記事としては，BBC(5 記事)，The Nation(5 記事)，Independent(5 記事)，New York Times(5 記事) に我々が正解データを人手で与えたものを用いた．ここでは，ウィンドウ幅は一文とした．

- 評価方法
  評価のための比較として，データ (a) に対しては「表層」: 表層マッチングのみ行なう場合，と「表層＋品詞」: 表層マッチングを行なったあとに品詞情報を用いたフィルタリングを行なう場合を比較した．データ (b) に対しては「表層＋品詞」のみ行なった．

- 正解候補イディオムの基準
  正解候補イディオムの基準としては，(1) テキスト中で実際に使われているイディオム，(2)(1) が “come to the point” の場合，“come to a point” のような冠詞が異なる（a と the の違い）イディオム，(3)(1) が “in spirits” の場合，“in spirit” のような語の単数形と複数形が異なるイディオム，(4)(1) が “come to the point” の場合，“to the point” のような別のイディオムの一部に含まれるイディオム，と設定した．図 5 に正解候補イディオムの位置付けを示す．



(A):正解イディオム（文中で使われているイディオム）
(B):正解候補イディオム（定義済み）
(C):翻訳者の理想の検索イディオム

図 5: 正解候補イディオムの位置付け

- 評価尺度
  本システムの評価尺度としては以下の尺度を利用した．翻訳支援の立場からは，テキスト中に出現するイディオムをとりこぼさないことが重要である．よって本システムでは再現率を重視し，高い再現

率を保ったまま，精度を向上することを目指す．

$$\text{精度} = \frac{\text{システムが検出した正解候補イディオム数}}{\text{システムが検出したイディオム数}}$$

$$\text{再現率} = \frac{\text{システムが検出した正解候補イディオム数}}{\text{総正解候補イディオム数}}$$

- 実験結果

表 2，3 に実験結果について示す．データ (a) は金平ら [3] で用いられている英語母語者によって作成されたイディオムの異形を含む文（100 文），データ (b) は実際の英語記事 (20 記事)，を対象にそれぞれ行なった結果である．データ (a) に対しては，品詞情報を用いたフィルタリングを行なうことで，表層マッチングのみによる検索よりも再現率をあまり落すことなく精度を向上させることができた．データ (b) に対しての実験結果は，データ (a) に比べ，再現率は高いが，精度は低い．この原因としては，(1) 実際の記事は 1 文が長く，誤ったイディオムを検索し易い，(2) フィルタリングの規則はデータ (a) の解析を基に作成しているので，実際の記事に用いる場合に不足がある，が挙げられる．

表 2: データ (a) に対するイディオム検索精度

	精度	再現率
表層	0.418 (218/521)	0.991 (218/220)
表層 ＋ 品詞	0.734 (213/290)	0.968 (213/220)

表 3: データ (b) に対するイディオム検索精度

	精度	再現率
表層 ＋ 品詞	0.528 (450/853)	0.978 (450/460)

4 考察

実験において誤ったイディオムの検索や正解イディオムの検索漏れの原因について述べる．原因としては，(1) 辞書に登録されているイディオムである，“make A of B” の “A, B” や，“have … in” の “…” 等を含むイディオムの処理，(2) 品詞情報を利用したフィルタリング規則の不足，(3) 入力文が “I graduated from high school.” の場合に検索される “from high”（意味：天から）のような，品詞情報を利用したフィルタリングルールでは削除できないイディオム，(4) イディオムの異形として句が挿入されている場合，がある．(1) に関しては，“A, B” や “…” にはどのような語句が入るのかを知る必要がある．(2) に関しては，フィルタリング規則を補足することで解決可能である．(3) に関しては，構文解析が必要となってくるので，我々が定めた解析レベルでは解決できない．(4) に関しては，句のパターンを調べる必要がある．

5 まとめ

本稿では，英語イディオムの自動検索システムの開発について述べた．“挿入” による異形をもつイディオムの検索は，品詞情報を用いることで検索精度を向上できることを明らかにした．またイディオム検索において，我々の解析レベルで解決できる限界についても明らかにした．今後はさらなる精度の向上と “置換” による異形をもつイディオムの検索を行なう必要がある．

謝辞

本研究の一部は，日本学術振興会科学研究費補助金基盤 (A)「翻訳者を支援するオンライン多言語レファレンス・ツールの構築」（課題番号 17200018）の支援を得て行われた．また,『グランドコンサイス英和辞典』のデータ利用につき,（株）三省堂に感謝する．

参考文献

[1] 影浦峡, 佐藤理史, 竹内孔一, 宇津呂武仁, 辻慶太, 戸田愼一, 小山照夫. 2006.「翻訳者支援のための言語レファレンス・ツール高度化方針」言語処理学会第 12 回年次大会発表論文集.

[2] 専門翻訳者・ボランティア翻訳者 6 名への最終著者による聞き取り.

[3] 金平昂，豊島実和，竹内孔一，影浦峡. 2006.「英語イディオムの異形を整理する」 言語処理学会第 12 回年次大会発表論文集.

[4] LogoVista, http://www.logovista.co.jp/product/product_x.html
THE 翻訳, http://cn.toshiba.co.jp/prod/hon_yaku/index_j.htm
ATLAS, http://software.fujitsu.com/jp/atlas/

[5] ジャン・マケーレブ，岩垣守彦. 2003.『英和イディオム完全対訳辞典』東京：朝日出版者.

[6] TreeTagger
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

[7] 三省堂編修所 編．グランドコンサイス英和辞典
http://www.sanseido.co.jp/publ/dicts/grand_con_eiwa.html