WordNet と同音異義語を利用した異形イディオム検索

蛭浜 康雄，金平 昴 †，平尾 一樹 †，竹内 孔一 †，阿辺川 武 ††，影浦 峡 ††
岡山大学工学部，† 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
†koichi@cl.cs.okayama-u.ac.jp，††{abekawa, kyo}@p.u-tokyo.ac.jp

1　はじめに

人手による翻訳作業において，文中に現れるイディオムに気が付くことは重要であるが，イディオムの数は多く実用的な数を暗記することはその言語に精通している必要があり，翻訳者にとって苦手とするところである．ところが，一部の例外を除き [2] 異形を含めた柔軟なイディオム検索システムの研究はあまり行われていなかった．そのため，人手による作業でも機械翻訳でもイディオムの扱いはボトルネックとなっており，我々は翻訳支援システムの一機能として，イディオム自動検索システムの構築を目指している．

本研究では，次章の中で述べる同音異義語による異形を新たに盛り込み，以前は同一視していた関連語の各語に対し，置換前の語との意味関係に基いた類似度を個々に計算する．この類似度を基本として，多様な異形のランク付けを同時に行う手法を提案し，実際のテキストデータに対して過剰な候補を絞り込むことができることを示す．

2　イディオムの異形

文中で現れるイディオムの異形については，様々な先行研究 [1, 4, 5, 6] が行われている．これらに加え翻訳者が実際に直面するイディオムの異形を我々は以下のように分類した．

2.1　異形の種類

(1) 主題化や受動化等，外部的文法操作による異形（“pull strings” → “these are the strings he’d happily pull”など）．
(2) イディオムの構成要素に直接関わる異形（“go halves” → “go exact halves”など）．
(3) 言葉遊びなどの生産的な異形（“screwed on right” → “screwed on wrong”など）．

これまでに行なっていた研究は (2)，(3) を対象にしており，本研究では主に (3) を広範囲で扱うことを目標にしている．以下にこれまでの (3) が対象としていた類義語，そして今回新たに研究対象とする同音異義語を含む異形について示す．

2.2　置換異形の形態

先行研究において，プロの編集者を含むネイティブスピーカーにイディオムの異形を作成して頂いたデータをもとに，異形のパターンを類型化した [9]．

先行研究 [9] から置換のタイプ分けはまず，シソーラスが表示する基本的な関係である反意語，同義・類義語，同列語の置換で名詞置換の約 74 % (241/327)，動詞置換の約 69 % (133/192)，形容詞置換の約 80 % (113/141)，副詞置換の約 65 % (26/40) を占めることがわかる．ここから，高品質のシソーラスを用いれば，置換による異形のかなりを扱えることが示唆される．この他に新たに外部資源を必要とするものとして,「その他」に含まれる，音が類似する単語へ置換する言葉遊び (“flower power” → “flour power” など) があり，本研究では主にこの 2 種類の置換形態を対象とする．

3　使用する言語資源

イディオム検索を行うにあたり，本研究では必要な言語資源として以下に示すデータを用いた．

3.1　イディオム・エントリー

本研究では，三省堂『グランドコンサイス英和辞典』[7] に記載されている抽出した約 25,000 のイディオムをイディオム辞書として使用している．この辞典を含め，通常の辞典にはイディオムの構成語について品詞情報は付与されていない．そのため本研究では構成語に品詞情報は保持させておらず，文中で一致した語の品詞を構成語の品詞としている．なお，我々が本研究を進める上で参考とした異形データに対するこの辞書のカバー率は 34%であった．今後，この数値を上げるために他の辞書と同時に用いることも検討される．

3.2　シソーラス

先行研究 [9] と同様に，置換の候補となる類義語を取得するためのデータとして，本研究では約 15 万語の名詞，動詞，形容詞，副詞が登録されている WordNet を用いている．しかし，前置詞については扱っていないため，独自に類義語を定義している [9]．置換による異形データから WordNet を用いて反意語，同義，同

列語の展開を行なったところ，正解データの約半分の異形をカバーできた．

類義語を取得する目的に加え本研究では，その構造を利用し，置換前の語と置換語の意味関係を測るために使用する．2 単語の非類似度は，比較する単語間の最短経路中に現れる関係から下の表 1 中の値を参照し合計を求め，その値に最短経路数の影響を与えている．

$$cost(word_1, word_2) = \{\frac{1}{SP_{num}} \sum_{p_i \in SPath} \sum_{rel_j \in p_i} weight(rel_j)\} + (1 - \frac{SP_{num}}{10})$$

$SPath$：最短経路
SP_{num}：最短経路の個数

表 1: 関係による重みづけの値

relation(rel)	weight(rel)
反意 (antonym)	1.1
上位・下位 (hypernym, hyponym)	1.2
部分 (holonym, meronym)	1.5
動詞グループ (verb-group)	1.0

3.3 発音辞書

同音異義語など発音を利用した言葉遊びでの異形をカバーするためには，英単語の発音辞書が必要である．本研究ではその対象として，『The CMU Pronouncing Dictionary』と呼ばれるフリーの英語発音辞書を使用した．辞書中には，単語の原形のみならず，複数形や過去分詞といった語形が変化したときの発音も個別にあり，計 125,000 種類以上の記載がある．そのため，文中で使われている語形の発音で同音異義語を得られ，言葉遊びを対象とする本研究にとってこの辞書は非常に有益なものとして採用した．

4 イディオム検索

4.1 検索アルゴリズム

イディオム検索システムの構成は図 1 のようになっており，その検索の流れを以下に示す．

1. 入力文を TreeTagger で解析し，各語の原形，品詞を取得
2. 各語の類義語・同音異義語をシソーラス，発音辞書から読み込む
3. イディオム辞書から各語とその関連語から始まるイディオムを得る
4. イディオムの構成語が文中の語もしくはその関連語と一致するか判定

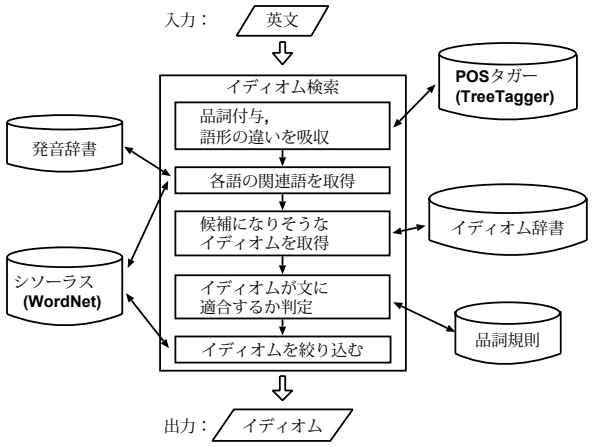

図 1: イディオム検索システム

5. 品詞規則を用いて不適切な挿入が行われているイディオムを削除
6. 絞り込み手法を用いて精度の高いイディオム候補のみを出力

まず始めに，入力文を POS タガー (TreeTagger) で各語の語形を吸収し，原形と品詞を取得する．これらを基にシソーラスから類義語を読み込む．例外として be 動詞は “being” と， 語形変化していない “be” の形のときのみ “doing”， “do” へ展開する．

類義語の展開と同時に，同音異義語の取得を行なう．発音辞書から入力文中の各語の発音情報を取り出し，同音異義語が存在すれば，それらを入力文中の語の関連語に追加する．

次に，文の先頭の語とその関連語で始まるイディオムをイディオム辞書から読み込む (先頭の語が冠詞など省略される語の場合もあるのでイディオムの先頭語のみで判断するのは最適とはいえないが，処理を高速にするために本研究ではこの方法を用いている)．この探索作業を文中に現れるすべての語について行う．図 2 にこれらの工程を行なった例を示す．

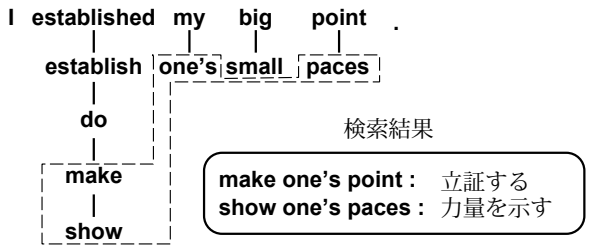

図 2: イディオム検索の様子

イディオム候補を取得し終えると，次はその取捨選択を行なう．まず挿入による異形の候補を厳選する．イディオムの構成語間に不適切な品詞の語が挿入されているものを品詞規則 [3] を用いて判定し，振り落とす．

最後に，主に関連語によって生じた過多な候補を取り除き，適切な候補のみを出力するため，絞込みを行う．この手法については次節で示す．

これらの工程を終え，最終的に残った候補を文中で使われている可能性が高いイディオムとして出力する．

4.2 候補の絞り込み手法

前節の途中で検出したイディオム候補は当然過多となるため，主に置換による異形に対する，より精度の高い候補のみを残す以下の絞り込み手法を提案する．

(a) イディオムの構成語が全て置換されているもの，もしくは 3 語以上から成り，置換されていない語が名詞，動詞，形容詞，副詞のいずれでもないものを削除

(b) 非置換イディオムと重なっている置換イディオムを削除．ただし，非置換イディオムを包含しているのであれば対象としない

(c) WordNet を用いて，置換前の語と置換後の語の意味関係を測り，意味が遠いものを削除する

(d) 各イディオム候補の適合値を計算し，ある候補が値がより良い候補と構成語が同じ位置を指す，もしくは包含されていればその候補を削除する．

(a) では置換による異形の候補として不適切なものを排除する．イディオムの構成後は全て置換されることは無く，置換されていない語があるとしても，イディオムの構成上で重要な名詞，動詞，形容詞，副詞は置換されずに残るため，非置換語がこれら以外の候補は削除対象とする (“have a seat” の異形として “take a seat” は可能だが，“take a stand” はあり得ない)．

(b) では，置換による異形が他に比べ出現頻度が低いことを利用し，非置換イディオムと部分的に重なっている置換イディオムを排除している．置換イディオムが非置換イディオムを包含している場合は，意味的にも包含していると考えられるので削除対象とはしない．

(c) で置換された語とその元の語の意味関係を WordNet の構造から測り，意味が離れた置換が行われた候補を排除している．WordNet の構造で，各語が属する synset 同士のノードの距離で意味関係を測定し，距離が 3 以上のものが対象となる．なおこの手法は意味関係を利用するため，同音異義語に対しては用いない．

(d) は，イディオムの構成語が示す入力文中の語が同じものは，より文中に適合するもののみを残すようにしている．以下の式を用いて非適合値 (uncmp : uncompatibility) を計算しており，置換された語の意味関係コストを合計したものを α 倍，挿入された単語総数の β 倍から構成語数を引いたものとしている．各定数については，$\alpha = 10$，$\beta = 10$ が最もより精度の高い結果となった．なお，前置詞の類語，同音異義語による置換のときの cost の値は 1 としている．

$$uncmp(idiom) = \alpha * \sum_{word_i \in idiom} cost(word_i, before(word_i)) + \beta * InsNum(idiom) - Num(idiom)$$

$before(word)$: $word$ の置換前の語
$Num(idiom)$: イディオムを構成している語数
$InsNum(idiom)$: 構成語間に挿入された語の総数

この式を異形データに適用したところ，バラツキはあるが最大の非適合値として 90 という数値を得た．本研究では再現率を優先するため，これを閾値とし，非適合値が閾値を超える候補も (d) の削除対象とする．

5 実データに対する実験

前章で述べたアルゴリズム・絞り込み手法の精度を測るため，以下の 2 種類のデータに対して実験を行った．

5.1 異形イディオムを含む文に対する実験

絞り込みの手法の効果を測定するため，挿入・置換による異形イディオムを含む文をそれぞれ 100 件ずつを対象にし実験を行った．この実験の対象データは 1 文と検出すべき正解イディオム 1 つの対という構成である．そのため本実験では，文を入力し，正解イディオムが出力候補に含まれていれば正解とする．

表 2: 挿入・置換の異形イディオムに対する実験

適用手法	挿入 100 件		置換 100 件	
	再現率	適合率	再現率	適合率
なし	0.93 (93/100)	0.005 (93/18616)	0.64 (64/100)	0.004 (64/16969)
(a)	0.93 (93/100)	0.015 (93/6059)	0.63 (63/100)	0.010 (64/6388)
(a),(b)	0.93 (93/100)	0.100 (93/934)	0.58 (58/100)	0.041 (58/1422)
(a)〜(c)	0.93 (93/100)	0.101 (93/919)	0.57 (57/100)	0.040 (57/1411)
(a)〜(d)	0.93 (93/100)	0.195 (93/476)	0.56 (56/100)	0.095 (56/587)

結果は表 2 に示すように，挿入 100 件に対して適合率は最良で 18%程度だが，再現率は 90%以上の数値を得ることができた．置換 100 件に対しては再現率は 60%と下がり，適合率に関しては 10%に満たなかった．

候補数は両方とも「なし」,「(a) のみ適用」では同程度だが，特に (b)，(d) が適用されたときに差が生じている．これは，手法 (d) で削除の対象となる「同じ入力文中の語を指す候補」が，挿入 100 件では (b) の段階で多く削られるためである．なぜなら挿入 100 件

では，挿入による異形を含む非置換候補が置換 100 件に比べ出現しやすく，非置換候補と置換候補の部分的な重なりが多く発生したためである．これに対し，置換 100 件ではこの削除対象が主に (d) で削除されることとなる．そのため，最終的に減少した規模は同程度だが，より (b) の効果が得られる挿入 100 件の方が候補数が少なくなった．

なお，出力された候補のうち非置換イディオムが挿入データでは 43%， 置換データでは 30%であり，本研究で置換による異形をカバーする手法を実装したが，出力の半数以上が置換による候補となった．これより，置換による候補は依然多数あり，適合率を上げるため，今後別に何らかの手法を用いる必要がある．

5.2 記事に対する実験

イディオム検索システムの精度を測定するため，実データに対する実験を行った．入力データは WEB 上から取得した BBC，Indie(The Independent)，NYT(The New York Times)， Nation (The Nation.) の 4 種類を 5 記事ずつを用いた．こちらは事前に人手により検出すべき正解イディオムを与えている．全記事中には置換による異形イディオムは含まれていなかったが，置換による過多な候補が出る量を測るために置換を含めた実験を行った．なお，この実験では絞り込みの手法を全て適用し，その実験結果を表 3 に示す．

表 3: WEB 上の新聞記事に対する実験

	置換無し		置換有り	
	再現率	適合率	再現率	適合率
BBC	1.00 (50/50)	0.40 (50/123)	1.00 (50/50)	0.08 (50/636)
Indie	0.96 (50/52)	0.34 (50/148)	0.96 (50/52)	0.09 (50/555)
NYT	0.85 (86/101)	0.38 (86/224)	0.85 (86/101)	0.08 (86/1045)
Nation	0.93 (187/201)	0.45 (187/414)	0.93 (187/201)	0.13 (187/1415)

挿入による異形に対して行った実験と同様に，正解イディオムに置換による異形が含まれないため再現率が 90%を超える結果となった．適合率に関しても，置換による異形を考慮した場合は同様に低い数値だった．置換なしとありを比較すると，適合率が 3〜5 分の 1 となっており，これに関しても前実験と同様に置換を考慮すると，数倍の過多な候補が生成される結果となった．

6 考察・まとめ

本研究では，文中に現れる置換による異形をカバーするため，WordNet による類義語の展開のみでなく，同音異義語による展開も行なった．先行研究とは実験の条件が違うため単純には比較できないが，絞りこみなど幾つかの手法を加えることで，再現率，適合率ともにより高くなっていた．適合率については，数値は低いものであるが，およそ 1 文に 5，6 個の候補となるので，翻訳を支援するレベルとしては負担にならないだろう．むしろ，結果の細かな判断は人により揺れるため [8]，過度に限定した候補のみを出力するより良いとも言える．今後の課題として，統語的な異形への対応，挿入による異形への精度の向上などがあり，より翻訳者を支援するために質を上げる予定である．

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」(研究代表者:影浦峡) （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』のデータ利用を許していただいた（株）三省堂に感謝する．

参考文献

[1] Baldwin, T.: Multiword Expressions, *Advanced course at the Australasian Language Technology Summer School* (2004).

[2] Carl, M. and Rascu, E. “A Dictionary Lookup Strategy for Translating Discontinuous Phrases” *EAMT-2006* (2006).

[3] 金平昂, 平尾一樹, 竹内孔一, 影浦峡: 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, 言語処理学会第 12 回年次大会発表論文集, pp. 711–714 (2006).

[4] Moon, R. 1998. *Fixed Expressions and Idioms in English.* Oxford: Clarendon Press.

[5] Nicolas, T. “Semantics of idiom modification,” In Everaert, et. al. eds. *Idioms: Structural and Psychological Perspectives.* Hillsdale: Lawrence Erlbaum Associates., 1995. p. 233-252 (1995).

[6] Numberg, G., Sag, I. and Wasow, Th. “Idioms,” *Language* 70(3), p. 491-538 (1994).

[7] 三省堂編集所 『グランドコンサイス英和辞典』 (2004).

[8] Sharoff, Serge and Babych, Bogdan and Hartley, Anthony “Using Comparable Corpora to Solve Problems Difficult for Human Translators” *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* pp. 739?746 (2006)

[9] 竹内孔一, 金平昂, 平尾一樹, 阿辺川武, 影浦峡: 置換・挿入を考慮した異形イディオム検索システムの構築, 言語処理学会第 13 回年次大会発表論文集, pp. 396–399 (2007).