2005－NL－170（9）
2005／11／21

論文データベースからのイディオム用例検索

難波 英嗣[1]　森下 智史[2]　相沢輝昭[1]

1 広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1

2 NEC マイクロシステム 〒211-0063 神奈川県川崎市中原区小杉町 1-403-53

E-mail: 1 {nanba, aizawa}@its.hiroshima-cu.ac.jp

あらまし 英語イディオムの用例を学術論文データベースから自動収集する手法を提案する．本研究では，特に，“regard 〜 as” のように，構成語が文中で不連続なイディオム(分離型イディオム)を対象にする．従来は，分離個所に含まれる単語数を制限して分離型イディオムの検索を行っていた．しかし，この方法では，分離個所に節を含むような用例を検索することができなかった．そこで本研究では，階層距離という尺度を定義し，構文レベルでの構成語の距離を測る手法を提案した．実験の結果，精度 0.862，再現率 0.726 が得られ，提案手法の有効性が確認された．

キーワード イディオム，論文，用例検索，構文解析

Searching Example Sentences of Idioms in a Research Paper Database

Hidetsugu NANBA[1]　Satoshi MORISHITA[2]　Teruaki AIZAWA[1]

1 Hiroshima City University　3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

2 NEC Micro Systems　1-403-53 Kosugi, Nakahara-ku, Kawasaki 211-0063 Japan

E-mail: 1 {nanba, aizawa}@its.hiroshima-cu.ac.jp

Abstract In this paper, we propose a method to search example sentences of English idioms from a research paper database. We focus on decomposable idioms such as “regard - as.” Traditionally, the decomposable idioms have been searched by limiting the maximum number of words between idiom particles. However, this method could not collect example sentences, in which clauses are inserted between idiom particles. We therefore devise a measure that calculates the distance between idiom particles on a parse tree, and use it for decomposable idiom search. We conducted an examination, and obtained the precision of 0.862 and the recall of 0.726.

Keyword idiom, research paper, example search, syntactic analysis

1. はじめに

英語論文を執筆する，とりわけ日本語で書いた論文を英語で書き直すにあたっては，訳語選択の問題がしばしば生じる．辞書にはさまざまな訳語が示されるが，そのうちどれが最も自分の目的に適っているかについて，限られた数の用例から判断するのは決して容易なことではない．そもそも，辞書の限られた紙面では，より幅広い層に対応するために，ある専門分野でよく使われる特殊な表現が掲載されなかったり，逆に学術論文では通常用いられない語句が，頻出語句として示されたりすることがある．また，分野によって異同があることも考えられるであろう．そこで本研究では特に英語のイディオムについて，学術論文データベースから用例を自動収集し，ユーザに提示するシステムの開発を行う．

我々は，これまでに Web 上の PDF や Postscript 形式の論文データを収集して，引

用論文データベース PRESRI を構築し,Web 上で公開している[6][1]．PRESRI は，現在，計算機科学，物理学，化学，天文学，電気工学等について学術論文約 78,000 件を収録している．これらの論文集合から特定分野の論文を収集するには，キーワード検索機能が利用できるが，この他，引用関係を利用した分類技術[7]を用いることも可能である．また，PRESRI は，各論文の出典情報も保持しているため，ある特定の論文誌や国際会議でよく使われるイディオムの用例を検索するといった目的にも利用できる．本研究では,PRESRI が収集した論文データを用いて，ユーザに特化したイディオム用例の検索および提示を目指し，その要素技術として,与えられた任意の文書集合から,イディオムの用例を自動検索するシステムを開発する．

本論文の構成は以下のとおりである．次節では，本研究で扱うイディオムについて述べる．3 節では，イディオムの検索手法を提案する．4 節では，提案手法の有効性を確認するために行った実験および結果について報告する．5 節では結論を述べる．

2. 本研究で扱うイディオム

複数の語から構成される表現は，一般に Multiword Expression (MWE)と呼ばれている．Baldwin[3]は，MWE を以下のように分類している．

1. **Lexicalized phrase**
   1. Fixed expression
      “ad hoc”のように完全な固定表現を構成するものであり,活用変化がなく,構成語の間に他の単語が入ることがない.
   2. Semi-fixed expression
      “kick the bucket”のように，文脈によっては，“kick”が“kicked”と変化する．このように，構成語の間に他の単語が入ることはないが，一部の構成語が変化することがある．複合名詞もこの分類に入る.
   3. Syntactically-flexible expression
      構成語の中に別の語句が入ることがある表現.
      (例)“write up”→“write the memo up”
2. **Institutionalized phrase**
   構文的，意味的には複合語であるが,統計的に見ると特異である句.
   (例) “kindle excitement (興奮する)”

これらの中で，本研究では構成語が分離して出現する Syntactically-flexible expression (以後，分離型イディオム)の用例検索を目指す．その理由は，構成語が連続して出現する場合，単純な文字列照合だけで用例検索が容易に実現できるからである．Semi-fixed expression のように構成語の語形が変化する場合も,形態素解析器(例えば[9])を使って単語を原形に変えておけば,文字列照合ができる．これに対し，構成語が分離している Syntactically-flexible expression は,何らかの制約を加える必要がある．そこで，次節では，分離型イディオムの用例を検索する際の制約について述べる．

3. 分離型イディオムの用例検索

3.1. 関連研究

Baldwin らは“hand in”のように「動詞＋不変化詞」で構成される表現を verb-particle constructions (VPC)と呼び，VPC の抽出に関する研究を行っている[1,2]. VPC は，2 節で述べた Syntactically-flexible expression の一種である．Baldwin らの研究では,「システムにはテキストのみを与え,その中からすべての VPC を抽出する」という問題設定になっているため,「システムにイディオムとテキストを与え，テキスト中からイディオムを含んだ文(用例)を探す」という本研究の設定とは異なる．しかし，Baldwin の抽出手法は本研究と関連性がある.

Baldwin らは，以下に述べる手法で VPC の抽出実験を行っている.

1. 構成語間の出現単語数(以後,表層距離)が 5 語以下のもの

[1] http://www.presri.com

2. 1＋分離区間が名詞, 前置詞, 副詞 chunk のもの
3. 2+chunk grammar を考慮したもの

ここで，表層距離が 5 語以下という条件が VPC の抽出に有効であることは, Baldwin らの行った実験で確認されているが，VPC 以外の Syntactically-flexible expression を対象にする場合には，必ずしもこの条件が有効であるという保証はない．なぜならば，VPC 以外のイディオムでは分離区間に節が挿入されることもあり, その多くの場合は，表層距離が 5 語を超える．本研究では，英語論文を書く上で，なるべく多様な用例をユーザに提示することが重要であると考えており，表層距離が 6 語以上の用例の検索も試みる．

3.2. 提案手法

前節でも述べたとおり，分離区間の単語数を制限する方法では，分離区間に句や節が挿入された場合の多くは，検索できないという問題がある．そこで本研究では，別の方法での用例検索を試みる．

以下に “share A with B (B と A を共有する)” というイディオムの用例を示す．

> But Mr. Foley predicted few economic policy changes ahead, commenting that Mr. Major **shares** a very similar view of the world **with** Mr. Lawson.

この例では，“share” と “with” の間の分離区間に 6 単語からなる名詞句があり，従来の手法では検索できない．ここでこの文の構文木を図 1 に示す．この図からわかるように，“share” と “with” を含むノードは，構文木上で見れば近距離であることがわかる. そこで, 本研究では文の構文情報に注目し，木構造における構成語間の距離を用例検索に用いる．

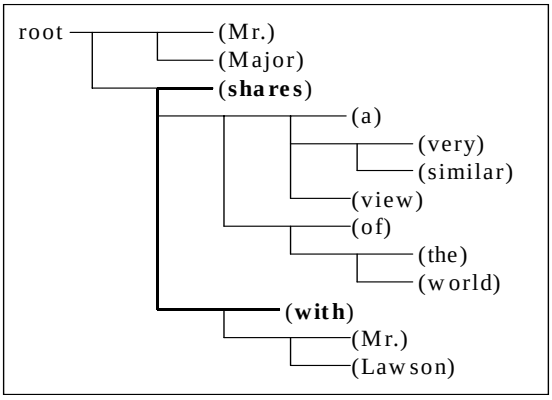

図 1 “share ～ with”の用例の構文木

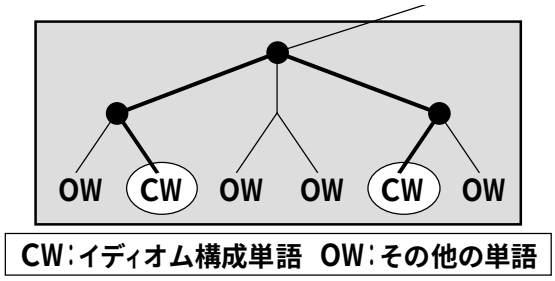

図 2 階層距離抽出例

本研究では，構文木上のノード間の距離(以後，階層距離)を，図 2 に示す例を用いて定義する．図 2 は，2 単語から構成される分離型のイディオムで，イディオムを構成する単語を CW で, それ以外のものを OW で表記している．階層距離は, 1 つ目の CW から 2 つ目の CW までの最短パスに含まれるノードの数と定義する．図 2 の場合，途中ノード(●)を 3 個所通過しているため，階層距離は 3 となる．本研究では，分離型イディオムを検索する際，閾値以下の階層距離内に CW が存在する場合，イディオムの用例の候補として検索する．

また，階層距離の他に，以下に述べる 2 種類の制約,「態の変換の禁止」「節の挿入」もあわせて考慮することで，より高精度な用例検索を目指す．

◆ **態の変換の禁止**

他動詞を含むイディオムは，態を変換して使うのはあまり一般的な用法ではないと考えられる．もし，あるイディオムが受動態で使うのが一般的であるならば，例えば

“be attributed to (〜のせいである)” のように，辞書の見出しにも受動態で表記されるからである．

この考え方の妥当性を調べるため,「コウビルド英英辞典」,「研究社 リーダーズ英和辞典＋プラス V2」,「研究社 新英和中辞典第 6 版」の 3 つの辞書を対象に，動詞型イディオム 21 個の用例で態の変換が行われているかどうか調べた. 結果を表 1 に示す.

表 1 辞書中で態の変換があった用例数

	用例数	態の変換があった用例数
コウビルド	17	2
リーダーズ	15	2
新英和中辞典	13	2

3 つの辞書には調査対象の動詞型イディオムに関する 45 個の用例があり, このうち態の変換があったのは 6 個(13%)であった. 十分な数での調査ではないので，この結果だけから断言することはできないが，動詞型イディオムの態の変換はあまり一般的ではないと言える．そこで，本研究では，動詞型イディオムの用例を検索する際，態が異なるものは検索対象から除外する.

◆ **節の挿入に関する制約**

分離型のイディオムでは，分離区間に節が挿入されることがある．本研究では，このような用例も検索対象とする．あるイディオムの分離区間で始まる節が，その分離区間内で終わっていれば，挿入節であると判断し，検索対象とする．逆に，その節が分離区間内で終わらなければ，検索対象から外す．例えば “the same A as B (B と同じ A)” に関するイディオムを検索する場合，図 3 の例では，分離区間内で始まる who 以下の節は，分離区間内で終わっていないため，検索対象から外す.

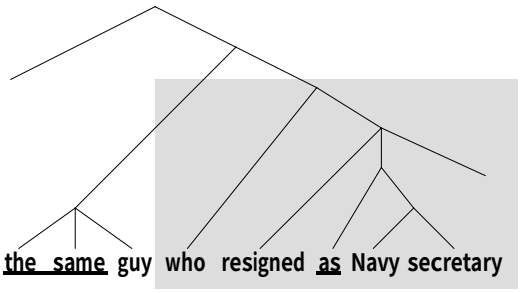

図 3 分離区間内で終わらない節の例

4. 評価実験

提案手法の有効性を確認するため，実験を行った.

4.1. 実験条件

● **実験方法**

本研究では，用例検索に構文解析器を利用するため，構文解析の精度が用例検索の精度に直接影響する．しかし，検索実験を行いエラー分析をする際，構文解析エラーの影響がどの程度であるのかが分からなければ，提案手法が有効であるかどうか判断できない．そこで，まず，理想的な(人手で付与した)構文情報を用いた場合にどの程度の検索精度が得られるかを調べておき，次に，構文解析器を使った場合の検索精度を調べる，という 2 段階で調査する．この調査が論文データを用いてできれば理想的であるが，我々の知る限り，構文情報を人手で付与した論文データは存在しない．そこで，代わりに Penn Treebank(PTB) [5]を用いて調査する．PTB は，新聞記事に人手で品詞および構文情報を付与したデータで, 約 74,000 文から構成されている. 本研究では，まず，PTB の構文情報に提案手法を適用し，理想的な構文解析が行われた場合の提案手法の有効性を調べる．次に，構文解析器による解析結果を用いて用例検索を行った場合の検索精度を調べる.

この他に，論文データを用いた実験も行う．この実験では，構文解析器の結果を使って検索をした場合の評価のみを行う．なお，実験には，PRESRI が収集した英語論文データ約 18,000,000 文を用いる.

● **比較手法**

以下の 3 種類の提案手法①～③で実験を行った.階層距離条件では,規則作成の際,最も検索精度の高かった値 4 を用いた.

また，提案手法との比較のため，2 種類のベースライン手法 i，ii でも実験した.ベースライン手法 i は，分離区間の単語数に制限のない検索手法であるため検索精度は低いが，正解の用例は漏れなく収集することができる.ベースライン手法 ii は，分離区間の単語数を制限する手法であり，今回は，規則作成の際最も検索精度の高かった値 3 を用いた.

提案手法：

① 階層距離条件（上限 4）
② ①＋態の変換の禁止
③ ②＋節の挿入に関する制約

ベースライン手法：

ｉ）単純マッチング（＝条件無）
ii）表層距離条件（上限 3）

● **正解データ**

検索に用いるイディオムは 53 個あり,このうち規則作成，評価共に用いられたものが 42 個，評価のみに用いられたものが 11 個である.使用するイディオムは全て 2～4 ブロックの分離型イディオムである.これらのイディオムは『科学技術英語表現辞典』[8]他 8 冊から選定した.

次に,上述の文書集合から 53 個の各イディオムに対して，以下の手順で正解用例を判定した.

1. LimaTK[9]を用い,イディオムと文書集合中の全ての単語を原形に変換する.
2. 単純マッチング法を用いて文を収集する.
3. 収集した文集合が入力したイディオムの用例であるかどうか，人手で判定する.

表 2 に，実験に用いたイディオム数，単純マッチング法により検索された文数，および正解数を示す.

表 2 実験用データ

		イディオム数	検索文数	正解数
PTB	規則作成用	42	662	429
	評価用	42	351	219
PRESRI		53	2466	1720

● **評価尺度**

上記のデータに対して提案手法を適用し，精度と再現率により評価する.

$$\text{精度}=\frac{\text{システムが検出した正解用例数}}{\text{システムが検出した用例候補数}}$$

$$\text{再現率}=\frac{\text{システムが検出した正解用例数}}{\text{総正解用例数}}$$

4.2. 実験結果

PTB の構文情報タグ付きデータを用いて用例検索を行った結果を表 3 に示す．表 3 からわかるように，提案手法はいずれもベースラインの精度を上回っている.

表 3 用例検索精度:PTB(人手)

		精度	再現率
ベースライン	I	0.624 (219/351)	1.000 (219/219)
	Ii	0.708 (155/219)	0.708 (155/219)
提案手法	A	0.796 (207/260)	0.945 (207/219)
	B	0.868 (204/235)	0.932 (204/219)
	C	0.868 (203/234)	0.927 (203/219)

次に，構文解析器を使って PTB のデータを解析し，その結果を用いて用例検索を行った時の結果を表 4 に示す.

人手で付与した構文情報を使った場合よりも構文解析器の結果を使った方が良くなっている理由は，対象文書に解析できない文が一定数あり，これらが提案手法の検索精度を下げる要因となる文の多くとたまたま一致していたためである.

表 4 用例検索精度:PTB(構文解析器)

		精度	再現率
ベースライン	i	0.624 (219/351)	1.000 (219/219)
	ii	0.708 (155/219)	0.708 (155/219)
提案手法	A	0.880 (205/233)	0.936 (205/219)
	B	0.887 (204/230)	0.932 (204/219)
	C	0.889 (201/226)	0.918 (201/219)

次に，論文集合を対象にした場合の検索結果を表 5 に示す．最も高い精度は，ベースライン ii を用いた場合であった．ただし，この手法は再現率が最も低い．

表 5 用例検索精度:PRESRI(構文解析器)

		精度	再現率
ベースライン	i	0.697 (1720/2466)	1.000 (1720/1720)
	ii	0.870 (1140/1311)	0.663 (1140/1720)
提案手法	A	0.841 (1303/1549)	0.758 (1303/1720)
	B	0.849 (1277/1505)	0.742 (1277/1720)
	C	0.862 (1248/1447)	0.726 (1248/1720)

4.3. 考察

● ベースライン ii と提案手法の比較

PTB を使った実験(表 3, 4)では，ベースライン ii と提案手法には 10%前後の精度の差があったにもかかわらず，PRESRI データを使った実験(表 5)ではほぼ同じ精度となっている．

その原因のひとつは構文解析器の解析精度と関連していると思われる．今回用いた構文解析器[4]は，PTB を用いて学習しているため，PTB を対象にした場合と比べ，PRESRI データを対象にした時の解析精度は低下すると推測される．

構文解析の用例検索精度への影響を調べるため，以下に述べる調査を行った．分離区間の表層距離が長いイディオムを提案手法で検出する場合，表層距離が短い場合よりも，構文解析結果の影響をより強く受けると推測される．そこで，分離区間の表層距離が 3 より大きいイディオムを対象に，PTB と PRESRI でどのくらいの精度で検索できるのか調べてみた．その結果，PTB では 97.6%(39/40)であったのに対し，PRESRI では 78.5%(325/414) と大きな開きがあることがわかった．

● 提案手法の有効性について

今回の実験結果を精度という点で見れば，単純な手法であるベースライン ii の方が，提案手法よりも優れているにが，この手法では，分離区間の表層距離が 4 以上のものはすべて検索対象から除外するため，今回比較した手法の中では再現率が最も低くなっている．特に，ユーザが英文を書く時になるべく多様な用例を提示する必要がある，という観点から考えると，精度だけでなく再現率もある程度重視する必要がある．さて，提案手法は，精度ではベースライン ii に及ばないものの，再現率では ii に優っている．そこで，両者を組みあせることで，お互いの欠点を補う検索が実現できるのではないかと考え，さらに調査を行った．

今回実験で用いているベースライン ii の表層距離と，提案手法 C の階層距離の閾値(それぞれ 3 と 4)は，PTB 規則作成用データから得られたものであるが，PRESRI でこれらの値を変化させた場合の再現率と精度を調べた．結果を図 4 に示す．図より，ベースライン ii は表層距離が 4 以上になると急激に精度が低下することがわかる．提案手法 C は，階層距離が 4 の時，最も解析精度が高くなり，4 より大きくなると低下する．ここで，表層距離が 1〜3 の時のベースライン ii の精度は提案手法 C の精度よりも高いが，表層距離が 3 よりも大きくなると逆転する．そこで，表層距離が n 未満の時にはベースライン ii で，表層距離が n より大きい時には提案手法 C で解析を行った． n の値を 2〜5 に変化させた時の結果(△)もあわせて図 4 に示している．図からわかるとおり，混合手法は，精度をほとんど低下させることなく高い再現率を得ることができている．F-measure で比較したところ，n が 4 の時の混合手法の F 値が最も大きいことがわかった(F 値 0.870，精度 0.832，再現率 0.911).

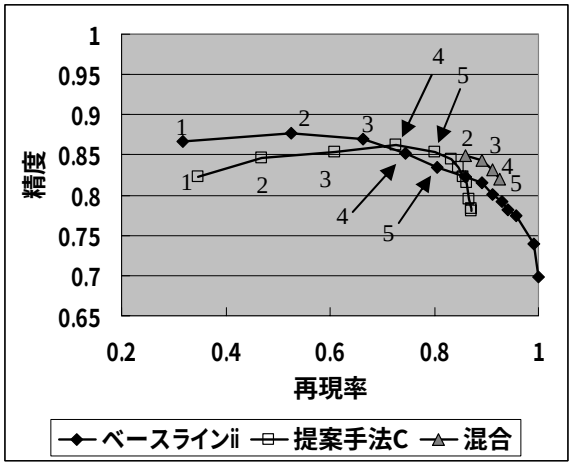

図 4 ベースライン ii, 提案手法 C, および両者を組み合わせた手法による検索精度 (PRESRI)

5. おわりに

本研究では，構文構造を考慮した分離型イディオムの用例検索手法を提案した．提案手法の有効性を確認するため，実験を行った．実験では，分離区間に含まれる単語数(表層距離)を制限するベースライン手法と比較した．実験の結果，提案手法では精度 0.862，再現率 0.726 が，ベースライン手法では精度 0.870, 再現率 0.663 が得られた．実験結果を分析したところ，表層距離が小さい場合はベースライン手法が，大きい場合は提案手法が優れていることがわかった．そこで，両者を組み合わせて実権行ったところ，表層距離が 4 未満の時にベースライン手法を，4 以上の時に提案手法を使って用例検索した時に，最も高い検索精度(精度 0.832，再現率 0.911)が得られた．

なお，現在，自然言語処理分野の論文を対象にしたイディオム用例検索システムを以下の URI で公開している．

http://www.nlp.its.hiroshima-cu.ac.jp/assist.html

参考文献

[1] T. Baldwin , A. Villavicencio , “Extracting the Unextractable: A Case Study on Verb-particles,” In Proceedings of the 6th Conference on Natural Language Learning 2002, pp. 98-104, 2002.

[2] T. Baldwin, C. Bannard, T. Tanaka, D. Widdows , “An Empirical Model of Multiword Expression Decomposability, ” In Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis , Acquisition and Treatment, pp.89-96 , 2003.

[3] T. Baldwin “Multiword Expressions,” Advanced course at the Australasian Language Technology Summer School, 2004.

[4] D.M. Bikel, “A Distributional Analysis of a Lexicalized Statistical Parsing Model, ” In Proceedings of the 2004 Conference on Empirical Methods in Natural language Processing, a Meeting of SIGDAT, pp.182-189, 2004.

[5] M. Marcus, G. Kim, M.A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz , B. Schasberger , “The Penn Treebank: Annotating Predicate Argument Structure, ”In Proceedings of the Human Language Technology Workshop, pp.114-119, 1994.

[6] H. Nanba, T. Abekawa, M. Okumura, S. Saito, “Bilingual PRESRI: Integration of Multiple Research Paper Databases, ” In Proceedings of the RIAO 2004, pp.195-211, 2004.

[7] 難波英嗣，神門典子，奥村学，“論文間の参照情報を考慮した関連論文の組織化,” 情報処理学会論文誌，Vol.42, No.11, pp.2640-2649, 2001.

[8] 富井篤 編，科学技術英語表現辞典，オーム社出版局，1995.

[9] 山下達雄，松本裕治，“言語に依存しない形態素解析処理の枠組，”自然言語処理，Vol.7，No.3，pp.39-56，2000.