JSiSE Research Report
vol.31,no.5(2017-1)

大規模コーパスを用いた言語処理に基づく英語複合名詞の学習支援

岩田　陽也[*1], 梅村　祥之[*1]
*1 広島工業大学　大学院工学系研究科

Assisting Education of English Noun Compounds by Corpus Processing

Yoya Iwata[*1], Yoshiyuki Umemura[*1]
*1 Hiroshima Institute of Technology Graduate School of Science and Technology

本稿では大規模コーパスを用いた英語複合名詞の学習支援方法を扱う．日本人英語学習者にとって複合名詞の使い方は難しいのに，複合名詞一覧などの学習教材がない．そこで，本研究ではコーパスから複合名詞を自動獲得する．手法には主に出現頻度を用いた．その結果，コーパスを用いての複合名詞の獲得が可能であることを確認し，誤抽出や特殊な語がわずかで，既存の辞書に記載されていない複合名詞を獲得出来た．

キーワード: 学習支援システム，言語処理，大規模コーパス，英語表現，複合名詞

1. はじめに

多くの日本人にとって英語で複合名詞を書くことは難しい．例えば,「学生用のアパート」と英語で書く場合, an apartment for students, または複合名詞を用いて a student apartment と書くことも出来る．複合名詞は物事を簡潔に表す事が出来るという特徴がある．例えば「実験室の空調の設備の」と表現すると「の」が何度も出現して長くて冗長な印象を受ける．その代わりに「実験空調設備」と表現すればすっきりとした表現となる．次に，複合名詞は生産性が高い． 生産性が高いとは，例えば「ネットの市民権」という言葉に対して「ネット市民権」という複合名詞を生産できることを意味する．したがって，複合名詞は辞書に載っていなくても自分で作れる点が特徴である． 文献(1)には「浴室用タオル掛けのデザイナーの養成」という言葉に対して bathroom towel rack designing training という複合名詞の実例が掲載されている．

以上の特徴により，複合名詞は辞書に載っていない表現でも自ら作成が可能である．しかし，日本人の英語学習者が英語で複合名詞を書く場合は単純に単語を直訳して，つなげれば良いわけではない．例えば，初心者は「男の赤ちゃん」という表現を英語で書く場合，直訳して male baby と書くかもしれない．しかし，正しくは baby boy である．更に言えば，初心者は baby boy の意味を「子供っぽい少年」と誤解するかもしれない．英語と日本語では「語順」やニュアンスの違いから「使う単語」も違う．このようなものは単純に語を並べれば通じるわけではない．

以上から考えると，英語学習者は，自ら複合名詞を生産する能力を獲得する必要がある．その際に，英語の文法書を参照すると，複合名詞の構成方法の説明が掲載されているものの，通常，複合名詞の例は数例しか掲載されておらず，複合名詞を生産する能力の獲得が困難である．

そこで，本研究では，大規模な言語コーパスを利用して，自然言語処理の技術を用い，複合名詞を自動獲得し，その中で学習に役に立ち，なおかつ辞書に載っていない複合名詞の選定を目的とする．

これまでに，自然言語処理を用いた複合名詞の研究

が行われている．文献[6]は日本語の複合名詞の内部の構造を分析する研究である．複合名詞と関連の深いものに固有表現がある[7]．固有表現とは，組織名，人名，地名などで，例えば，「福島第一原発」等である．本研究で用いる手法は，基本的に固有表現抽出と同じ統計的言語処理を用いた手法である．

2. 複合名詞の自動獲得の予備検討

2.1 方法の概略

英文の大規模コーパスである英語版 Wikipedia[3]を用いて複合名詞の出現頻度を計測し，その値に基づいて，コーパス中にある程度出現する複合名詞を獲得する．

コーパス中で複合名詞の出現頻度を調べるにあたって，時間節約のため，対象となる複合名詞のリストをあらかじめ作成する．複合名詞のリストには，英語辞書から初級単語を用いることにする．英語辞書から初級レベルの名詞を数百語選定する．今回は基礎検討として初級単語 2 語からなる複合名詞を検討対象とする．そのために，初級単語 2 語を組み合わせ，複合名詞を機械的に数十万語生成する．その中には，複合名詞として成立しない，またコーパス内に出現しないものも多々含まれている．それらはコーパス中の出現頻度の統計量などで省く．

このようにして獲得された複合名詞を用いて，様々な条件で，どのような語が得られるのかを観察する．

2.2 獲得方法 Step1

まず，複合名詞の獲得を行うために，単語辞書「英辞郎第三版[2]」の初級単語にあたるレベル 1 及び 2 の中に含まれている単語 2,000 語を取る．名詞には一般名詞と固有名詞の 2 種類がある．基礎検討にふさわしく，より一般的な語として一般名詞を使う．調べた結果，一般名詞は 462 語含まれていた．

次に，462 語の一般名詞から 2 語を取り出して繋げた複合名詞を作成する．462×461 = 212,982 個の複合名詞となる．これら 212,982 個の複合名詞を含む文を英語版 Wikipedia 内で検索し，文の出現頻度を計る．以下，出現頻度は文内に複数回出現しても 1 回と数える．

Wikipedia は文章の質が高く，2015 年 12 月 1 日版で全体では約 30 億ワードを含むが，Step1 に限り，計算時間短縮のため，その一部である約 1 億 4 千万ワードで名詞 2 語からなる複合名詞を検索する．その結果，出現頻度が 1 以上のものが 4,187 語得られた．この 4,187 語を以下，様々な条件の下で絞り込む．

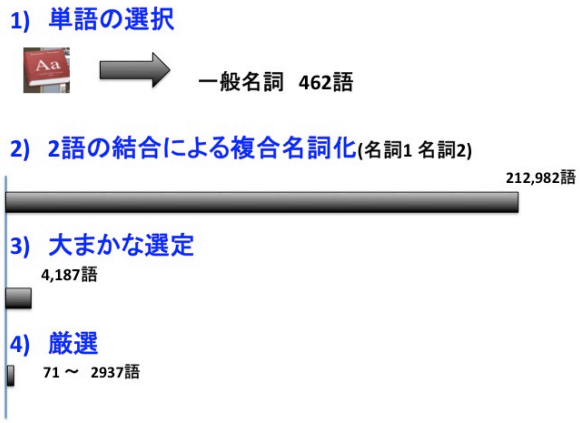

図 1 複合名詞の自動獲得の流れ

2.3 獲得方法 Step2

始めに概略を述べる．前節で得られた 4,187 個の複合名詞のうち，コーパス内に多く出現するものを選定する．選定条件は，複合名詞を構成する 2 語が同じ文中に現れる頻度の指標である共起頻度(詳細後述)と，複合名詞を構成する 2 語の共起しやすさに関する指標であるダイス係数(定義後述)から設定する．これらの 2 つの条件にそれぞれ閾値を設けて複合名詞を選定する．なお，本節以降は約 30 億ワードを含むコーパス全体で検索を行う．

以下，詳細を述べる． 4,187 個の複合名詞の共起頻度を計測する．その際，コーパスは英語版 Wikipedia 全文を用いる．また，検索にあたって，複合名詞の前後のいずれかに名詞があれば頻度に含めない．その際，英文形態素解析ツールの OpenNLP[4]を用いて，文中の単語の品詞の同定を行う．これは 3 語以上の名詞から成る複合名詞を除外するためである．

次に，複合名詞を構成する 2 つの名詞の共起しやすさの指標であるダイス係数を計測する．今回使用する複合名詞は「単語 1　単語 2」という形式になっている．「名詞 1」の出現頻度を freq(*名詞 1*)，「名詞 2」の出現頻度を freq(*名詞 2*)，「名詞 1+名詞 2」の出現頻度を freq(*名詞 1，名詞 2*)としたとき，ダイス係数 d

は次式で与えられる.

$$d = \frac{2freq(名詞1, 名詞2)}{freq(名詞1) + freq(名詞2)}$$

図 2 ダイス係数の数式

選定のための閾値を設定するにあたって，複合名詞の出現頻度とダイス係数の双方の常用対数値のヒストグラムを求める．結果をそれぞれ図 2，図 3 に示す．このグラフから，出現頻度の閾値として常用対数値 2.5，2.0，1.5，1.0，0.5，0.1，0.05，0.01，の 8 通りを設定し，ダイス係数の閾値として，常用対数値 −2.5，−3.0，−3.5，−4.0，−4.5，−5.0，−10.0，−50.0，−100.0，の 9 通りを設定する．両者の組み合わせ 29 通りを用いて複合名詞を抽出する．結果は，71 語から 2,937 語が得られた.

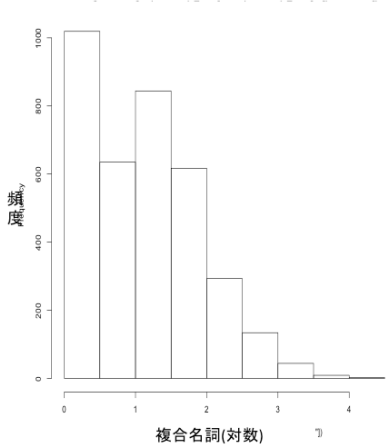

図 3　複合名詞の対数

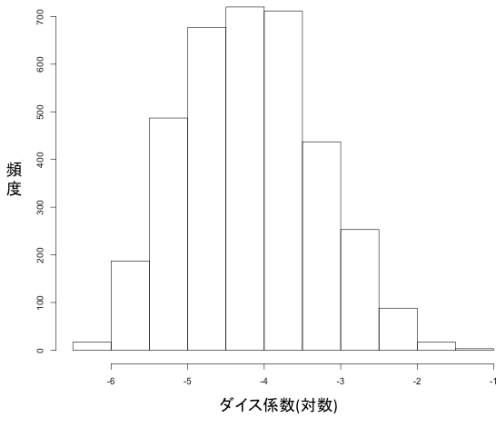

図 4　ダイス係数の対数

2.4　予備検討の結果及び考察

先の 29 通りのうち，代表的に，1 番厳選された条件，中間，1 番緩い条件，の 3 つの条件で得られた語の例を記す．これらの条件の結果を下に学習に適切な語の獲得の条件を考察する.

選定条件 1)

・複合名詞の頻度(常用対数) 2.5 以上

・　ダイス係数(常用対数) −2.5 以上

獲得語数:71

例:

〈名詞 1〉	〈名詞 2〉
car	accident (以下に用例を示す)
college	baseball
computer	science
opera	singer
tennis	player

選定条件 2)

・複合名詞の頻度(常用対数) 1.5 以上

・ダイス係数(常用対数) −4.0 以上

獲得語数:1,008

例:

〈名詞 1〉	〈名詞 2〉
child	actor
action	agency
campus	area
education	department
platform	height (以下に用例を示す)

選定条件 3)

・複合名詞の頻度(常用対数) 0.01 以上

・ダイス係数(常用対数) −50.0 以上

獲得語数:2,937

例:

〈名詞 1〉	〈名詞 2〉
speech	april
morning	zoo
meat	locker (以下に用例を示す)
century	children
day	building

以上は単語のみを表示している．それらの単語を出現した用例を以下に示す．

条件 1)

〈名詞 1〉　〈名詞 2〉

Car　accident

On 27 October 1998, in Vladivostok, Russia, Douglas Kent, the American Consul General to Russia, was involved in a car accident that left a young man, Alexander Kashin, disabled.

条件 2)

〈名詞 1〉　〈名詞 2〉

platform　height

A large number of stations had to be upgraded; platforms were extended to a length of to allow for three-unit trains, and the platform height was raised to .″

条件 3)

〈名詞 1〉　〈名詞 2〉

meat　locker

His pyromania continued and he was forced to run away after locking a schoolmate in his house and setting it on fire, after the boy locked Rory in a meat locker during a field trip.

予備検討を通じて，以下のことが確認出来た．個人のブログのような正しくない表現が多くでる Web ではなく，質の高い文章が書かれている Wikipedia を用いて，選定条件 1 で厳選した結果，日常で使われる複合名詞が多く出現した．しかし，これは既に日本人学習者が知っているものや既存の辞書に載っているものなど学習の必要が無いものである．対照的に，選定条件 3 のように緩い条件では，日本人が目にすることが少ない奇異で，辞書に載っていない単語が得られた．奇異単語もコーパス内の使用例を調べると英語圏で一般的に使用される語であった．このように学習する必要のある単語が多く獲得出来た．誤抽出がわずかに見られた．以下に，複合名詞と用例を載せる．

例）century movement

″Quakers were heavily involved in the 19th century movement for women's rights in America; the landmark 1848 Seneca Falls Declaration was in large part the work of Quaker women, and has numerous Quaker signatories, well out of proportion to the number of Quakers in American society at large.

2 語を超える長さの複合名詞の一部分が抽出された．誤抽出は，日付に関する表現のものであった．

2.5 予備検討の結論

1) 信頼の置ける複合名詞の獲得に Wikipedia を利用出来る

2) Wikipedia で獲得出来る複合名詞の中に，学習に役に立つものがいくつも含まれている

3. 本実験

3.1 目的

前章での抽出結果を踏まえて，機械的に獲得した複合名詞を対象に，次の 3 点を分析する．

1）辞書に掲載されている複合名詞及び掲載されていない複合名詞の割合を調べる

2）誤抽出の割合を調べる

3）得られた複合名詞が狭いコミュニティで使われる特殊なものとそうでないものの割合を調べる

3.2 方法

予備検討においては，2 つの指標を用いて，どのような語が得られるかを調べた．本実験ではそれ以上の検討を行わず，出現頻度のみを使う．選定条件を予備検討での条件 3 より少し少ない獲得数となる，「共起頻度が 10 以上」のみに設定する．その結果，2,008 個の複合名詞が抽出された．今回は基礎検討として，約 50%にあたる 1,000 語を調査対象とする．この 1,000 語のうち，誤抽出の割合及び，辞書に掲載されていない割合，及び特殊でない複合名詞の割合を調査する．前節で述べた目的，1)，2)，3)，に対して，作業者が 1)及び 2)では 1 名，3)では 2 名で確認作業を行った．作業者の英語力は，共に，TOEIC スコアが 600 点以上である．調査の方法を以下のフローチャートで示す．

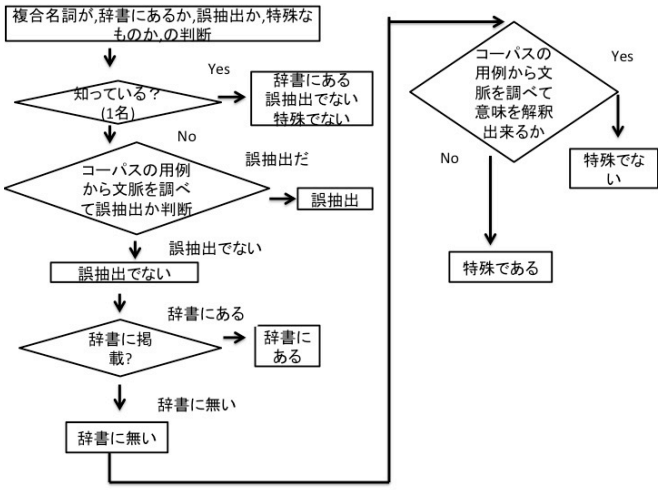

図 5　調査作業フローチャート

調査の結果，137 語の複合名詞が，誤抽出ではなく，辞書に掲載されておらず，特殊なものではないことが確認された.

3.3　構成率の調査結果

調査対象の 1,000 語に対し，上記のフローチャートに沿った構成率の調査を行った．その結果，全体の約 14%のあたる 137 語が辞書に記載されていないものだった．割合を下記の図に記す．また，得られた複合名詞は辞書には載っていないが，英語の母語話者が実際に使用しているものである.

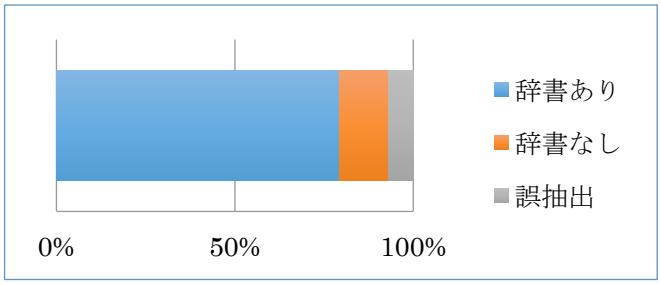

図 6　誤抽出，辞書に記載されている/いないの割合

以下に，本実験で得られた，辞書に記載されていない複合名詞とその文の例を載せる.

例 1)　camera car

For example, a camera car may drive along streets or roads while photographing the changing scene behind it.

例 2) attention saying

Then the other boy called my attention saying the same as the other had said.

例 3) system president

In March 2007, Lancaster announced that he would retire as system president in the spring of 2008.

4.　考察

Wikipedia の文に関して大まかに印象を評価したところ，質が高いことが分かった．獲得された 2,008 語中の調査対象の 1,000 語のうち，誤抽出は全体のわずか約 0.7% にあたる 69 語であった．全体の約 14%にあたる 137 語の既存の辞書に掲載されていない複合名詞が得られた．以上のことから，Wikipedia により，複合名詞の獲得が可能であり，学習に有用な複合名詞を含むことが分かった.

今後の展開としては，単語辞書と同じく数万規模の複合名詞辞書を作ることを考えると，今回の手法をそのまま適用し，英辞郎の初級単語だけでなく，レベルを上げて，更に多くの単語を使用すれば，数万規模の複合名詞辞書を自動獲得出来る見通しを得た.

5.　まとめ

今回の研究で得られた成果は以下の通りである

1)　Wikipedia で複合名詞の獲得が可能であることを確認した

2)　既存の辞書に記載されていない複合名詞を数多く獲得出来た

3)　誤抽出や狭いコミュニティで使われる特殊なものもわずかであった.

この手法を使えば，数万規模の複合名詞辞書を自動構築出来る見通しを得た.

参考文献

(1) 大津由紀雄，今西典子，池内正幸，水光雅則：“言語研究入門”，研究社，2002 年 5 月 1 日

(2) Electronic Dictionary Project 監修：英辞郎第三版，アルク(2007).

(3) https://en.wikipedia.org/wiki/Main_Page
2015 年 12 月 1 日版

(4) https://opennlp.apache.org/announcement/release-160.html

(5) http://ejje.weblio.jp/
weblio 英和辞典・和英辞典

(6) 小林義行, 徳永健伸, 田中穂積: “名詞間の意味的共起情報を用いた複合名詞の解析”, 自然言語処理, Vol.3,No.1(1996).

(7) 山田寛康, 工藤拓, 松本裕治: “Support Vector Machine を用いた日本語固有表現抽出”, 情報処理学会論文誌, Vol.43,
No.1,pp.44-53(2002).