

用例・規則融合型の複合名詞の日英翻訳方式

孟 莉*

川辺 諭†

宮崎 正弘*

* 新潟大学大学院自然科学研究科 † 科学技術振興機構

1 はじめに

日本語では、名詞や名詞相当語の接辞がいくつも連続することにより複合名詞が数限りなく作り出される。その対訳である英語にも様々な表現があり、そのすべてをあらかじめ対訳辞書に収録しておくことは、困難である。そのため、従来提案されている複合名詞の日英翻訳においては、頻出の複合名詞の対訳パターンを辞書に収録する方法が多い。また、特定の分野において非対訳コーパスを用いた複合名詞の翻訳法 [1][2] が提案されているが、長い複合名詞に対して十分な解析精度は得られていなかった。

本稿で提案する手法では、2名詞で構成される日本語複合名詞に関して約1万の日英名詞句の対訳用例データを準備し、入力された日本語名詞句と意味的に類似性の高い対訳用例を用いて、出力する英語名詞句の統語構造(訳型)を推測する。また、適切な訳語を選択するために、すべての英語名詞句の出力候補に関して検索エンジン google を用いてウェブページ上での頻度を確認し、出力される英語名詞句の確らしさを確認する。簡単な評価実験の結果を元に、本手法の有効性について論じる。

2 日英名詞句対訳用例データ

2.1 日本語品詞コード

日本語複合語を構成する語に関して宮崎 [3] の分類を元に、日本語名詞を図1のように細分化し、表1の品詞分類を準備した。

表1: 品詞と品詞コード

品詞	品詞コード
一般名詞	1100
動作名詞	1210-1230
状態名詞	1300-1600
数詞	1710-1760
副詞型名詞	1810-1830
固有名詞	1900-19F0
接頭辞	6110-6190
接尾辞	7110-79F0

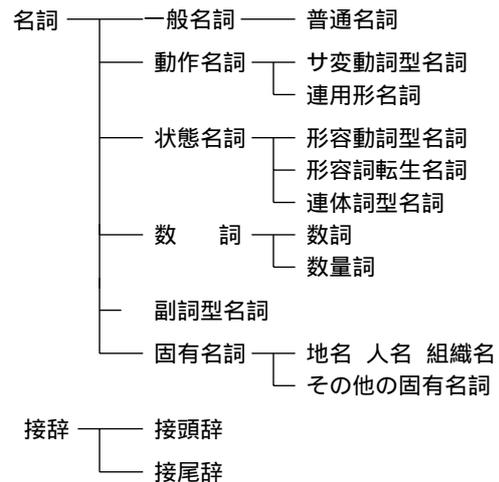


図1: 日本語名詞の細分化

2.2 日本語複合名詞データ

日英対訳用例 9690 件の日本語複合名詞に関して、表1の品詞分類を元に、前方・後方語の品詞の組み合わせで分類した。それぞれの件数を表2に示す。

表2: 日本語複合名詞の品詞構成による分類

組合せ	件数	組合せ	件数
一般名詞+一般名詞	1933	動作名詞+一般名詞	1173
一般名詞+動作名詞	1119	動作名詞+動作名詞	438
一般名詞+状態名詞	93	動作名詞+状態名詞	17
一般名詞+接尾辞	1609	動作名詞+接尾辞	1370
状態名詞+一般名詞	410	接頭辞+一般名詞	374
状態名詞+動作名詞	186	接頭辞+動作名詞	107
状態名詞+状態名詞	18	接頭辞+状態名詞	15
状態名詞+接尾辞	133	接頭辞+固有名詞	26

それぞれの組み合わせに関して、日本語複合名詞の例を表3に示す。

2.3 英語名詞句の形式

日本語複合名詞が連続した2名詞で構成されているが、英語名詞句は名詞に加えて、形容詞、前置詞など

表 3: 複合名詞の例

前方 \ 後方	一般名詞 1100,1810 ~ 1830	動作名詞 1210 ~ 1230	状態名詞 1300 ~ 1600	接尾辞 7100 ~ 79F0
一般名詞 1100 1810 ~ 1830	医療機関 宇宙科学 経済政策	価格操作 化学反応 石油輸出	技術標準 言語明晰 風景絶佳	工業国 価格差 基準点
動作名詞 1210 ~ 1230	観測衛星 運営資金	委託加工 選挙妨害	生産過剰 報道迅速	選挙戦 練習場
状態名詞 1300 ~ 1600	基礎体温 異常気象	自動振替 予備試験	軽快敏活 精鋭剽悍	小型車 専門医
接頭辞 6110 ~ 6190	前 総理 大 企業	仮 発注 密 結合	半 透明 高 精細	— —

表 4: 英語名詞句の訳型による分類

訳型	件数
N1.adj N2	3498
N	2184
N1 N2	1881
N2 prep N1	883
N1.ing N2	312
N1 N2.pl	273
N2 prep N2.pl	228
N2.pl prep N1	108
N1.pl N2	106
N1.s N2	94
N1 N2.ing	81
N1 N2.ing	24
N1.pl N2.pl	18

他の品詞を用いて表現される。英語名詞句は形式上以下の 13 種類の訳型に分類した。

- 直訳型 1 : N1 N2
日本語と英語がそのまま対応している場合
(「大学, 病院」; a university hospital)
- 直訳型 2 : N1 N2.pl
直訳型で、英語後方単語が複数形の場合
(「化学, 製品」; chemical products)
- 直訳型 3 : N1.pl N2
直訳型で、英語前方単語が複数形の場合
(「規格, 協会」; standards organization)
- 直訳型 4 : N1.pl N2.pl
直訳型で、英語の前方・後方単語両方が複数形の場合
(「貯蓄, 銀行」; savings banks)
- 前置詞型 1 : N2 prep N1
前置詞を介し、前方・後方名詞が入れ替わる場合
(「会社, 社長」; the president of a company)
- 前置詞型 2 : N2 prep N1.pl
前置詞型で、英語後方単語が複数形の場合
(「学生, 部長」; a dean of students)
- 前置詞型 3 : N2.pl prep N1
前置詞型で、英語前方単語が複数形の場合
(「政府, 首脳」; the leaders of the government)
- 前置詞型 4 : N2.pl prep N1.pl
前置詞型で、英語の前方・後方単語両方が複数形の場合
(「経済, 原論」; the principles of the economics)
- 形容詞型 : N1.adj N2
前方名詞が形容詞となって後方を修飾する場合
(「市長, 選挙」; mayoral election)
- 所有格型 : N1.s N2
前方名詞が所有格となる場合
(「社長, 室」; the president's office)
- 分詞型 1 : N1.ing N2
前方名詞が現在分詞となる場合
(「建築, 技術」; building techniques)
- 分詞型 2 : N1 N2.ing
後方名詞が現在分詞となる場合
(「価格, 操作」; price fixing)
- 接辞型 : N
接辞が修飾する名詞に接続され 1 単語となる場合
(「無, 着陸」; nonstop)

日英対訳用例中の英語名詞句 9690 件を、13 の訳型に分類した件数を表 4 に示す。

3 複合名詞の翻訳

試作中の日英複合名詞翻訳部の概要を図 2 に示す。

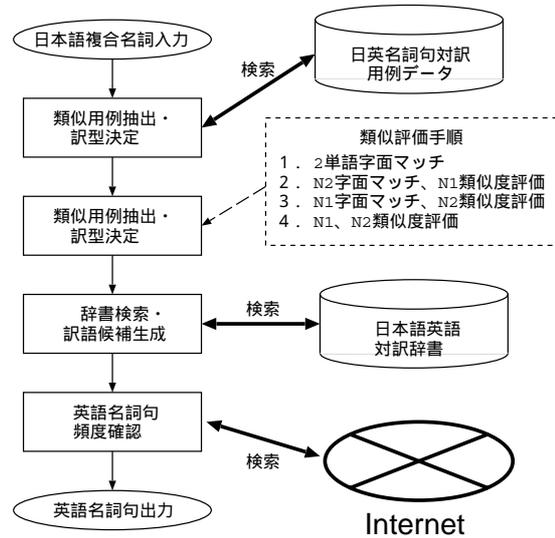


図 2: 試作中の日英複合名詞翻訳部

翻訳処理の流れは以下の通り。

- 日本語複合名詞の前方・後方語の字面と意味カテゴリーを手がかりに、意味的に類似性の高い日英名詞句対訳用例データを検索し、出力英語名詞句の訳型を決定する
- 日本語に対応する英語訳語の候補を訳型に応じて検索し、出力文の候補を生成する
- それぞれの候補に関して検索エンジン google を用いてウェブ上でのヒット件数を調査し、出力英語名詞句を絞り込む

3.1 訳型の決定

日英名詞句対訳用例データ中から、入力された日本語複合名詞に関して意味的に似通ったものを抽出し、出力する英語名詞句の訳型を決定する。単語類似度の計算は以下の手順で行う。

1. 前方・後方語字面一致

用例中に字面が完全に一致する日本語複合名詞がある場合は、対訳英語名詞句を出力とする

2. 後方語字面一致・前方語類似度算出

用例中に入力された日本語複合名詞と後方語が一致し、前方語の品詞が同じ候補がある場合は、各候補に関して前方語の単語類似度を算出し、結果が最も高い候補の訳型を選ぶ

3. 前方語字面一致・後方語類似度算出

2.とは逆に、前方語が一致し後方語の品詞が同じ用例に関して後方語の単語類似度を算出し、結果が最も高い候補の訳型を選ぶ

4. 前方・後方語類似度算出

入力された日本語複合名詞の前方語・後方語の品詞が同じ用例に関して、前方語・後方語双方の単語類似度を算出し、結果が最も高い候補の訳型を選ぶ

複合名詞は後方語が主名詞となることが多く、複合名詞全体は後方語の意味的・統語的性質を受け継ぐことがほとんどである。そのためここでは、後方語が一致する用例を重点的に採用している。

単語類似度の判定には、名詞意味属性体系 [4] を用いる。図 3 は一般名詞意味体系の一部である。

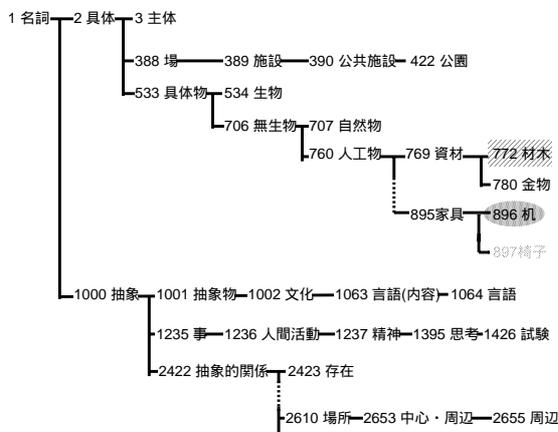


図 3: 一般名詞意味体系の一部

単語類似度は以下の式で算出する [5]。

$$s = \frac{d_c \times 2}{d_i + d_j} \quad (1 \geq s \geq 0)$$

ここで d_i 、 d_j は計算対象となる 2 ノードのシソーラス上での深さ、 d_c は共通ノードの深さである。

以下は単語類似度の算出の例である。

「机」と「材木」、「机」と「椅子」の類似度を計算シソーラス上

[机, 896, 895, 894, 893, 760, 706, 533, 2, 1]

[材木, 772, 771, 769, 760, 706, 533, 2, 1]

[椅子, 897, 895, 894, 893, 760, 706, 533, 2, 1]

「机」の深さは 8

「材木」の深さは 7

「椅子」の深さは 8

「机」と「材木」との共通ノードの深さは 4

類似度 : $s = 4 \times 2 / 8 + 7 = 0.533$

「机」と「椅子」との共通ノードの深さは 7

類似度 : $s = 7 \times 2 / 8 + 8 = 0.875$

3.2 訳語検索処理

訳語検索部では、日本語複合名詞の前方語・後方語に関して、日英対訳辞書を以下の手法で検索する。

- 一般名詞 : 名詞辞書を検索
- 動作名詞 : 語尾に「する」をつけて動詞辞書を検索
- 状態名詞 : 語尾に「だ」をつけて形容詞辞書を検索

次に、訳型に応じて以下の形態素調整を行ったものを検索し、すべてを訳語候補とする。

- N : 名詞、名詞複数形、形容詞名詞派生形
- N.pl : 名詞複数形
- N.ing : 動詞現在分詞形
- N.adj : 形容詞、名詞形容詞派生語、動詞現在分詞形、動詞過去分詞形、動詞形容詞派生語

対応する語形がないために形態素調整に失敗した場合は、N と N.adj の和集合を訳語候補とする。

3.3 英語名詞句絞り込み

訳語候補を組み合わせることで生成されるすべての英語名詞句候補に関して、検索エンジン google を用いてウェブ上での出現頻度 (ヒット件数) を調査し、生成された英語表現が妥当であるかどうかを判断する。以下は試作中の日英複合名詞翻訳部の実行例である。

```
% ./makeNp2.rb 開発 計画
訳型計算中 : N1 N2
頻度計算中 :
development plan : 1540000
development projection : 572
development project : 1930000
development scheme : 65200
development schedule : 43400
exploitation plan : 4490
exploitation projection : 11
exploitation project : 1230
exploitation scheme : 249
exploitation schedule : 33
result : development project
```

類似度計算により訳型として直訳型 N1-N2 が選択され、訳語検索処理で得られた 10 通りの英語名詞句候補を google を用いて検索することで、“development project” という英語名詞句出力が得られている。

4 実験と検討

試作中のシステムに関して評価実験を行った。

4.1 評価実験の結果

電子化辞書中の日英複合名詞対訳対 300 に対して行った評価実験の結果を表 5 に示す。

表 5: 評価実験の結果

出力結果	件数
正解	210
訳語が異なるが、意味的にほぼ同等	22
訳型、訳語選択の失敗	68
合計	300

正解率は 77.3%であった。

4.2 問題点：訳型選択誤り

訳型の選択を誤ったために正解が得られない場合がある。

```
% makeNp2.rb 職業 訓練
訳型計算中 : adj
頻度計算中 :
occupational training : 46100
occupational trainings : 43
occupying training : 27
occupying trainings : 0
occupied training : 59
occupied trainings : 1
result : occupational training
```

この例では “occupational training” が解として出力されているが、より適切な解は “job training” であり、こちらは google での検索件数は約 201 万である。訳型として形容詞型を選択したために、“job training”

が得られていない。この問題点を解決するには、日英名詞句対訳用例を拡充する必要がある。

4.3 問題点：訳出困難な語

訳語検索に失敗する場合がある。

```
% makeNp2.rb 割烹 料理
訳型計算中 : adj
頻度計算中 :
.. 訳語がありません
```

この例の正解は “Japanese cooking” である。このような訳出困難な語は、あらかじめ日英名詞句対訳用例として準備しておく必要がある。

5 まとめ

本稿では、日本語複合名詞の日英翻訳に関して、2 名詞相当語によって構成される日本語複合名詞に関して日英名詞句対訳用例データを用意し、意味的に類似度の高い用例から出力英語名詞句の訳型を決定し、訳語候補に関してウェブ上の出現頻度を比較することで、出力英語名詞句を得る手法を提案し、小規模の評価実験を行った。

今後の課題として、訳語選択の精度を向上するために日英名詞句対訳用例のデータ数を増やし、より正確な英語名詞句を生成することがあげられる。

参考文献

- [1] イラム, 大竹, 増山, 山本: 非対訳コーパスを用いた日本語複合名詞の英訳語推定, 情報処理学会研究報告, NL133-2 (1999.9)
- [2] 田中, 松尾: 対訳関係のないコーパスからの複合名詞対訳表現の獲得, 電子情報通信学会論文誌 Vol.J84, No.12, pp.2605-2614(2001)
- [3] 宮崎, 白井, 池原: 言語過程説に基づく日本語品詞の体系化とその効用, 自然言語処理, Vol.2, No.3, pp.3-25(1995).
- [4] 池原, 宮崎, 白井, 横尾, 中岩, 小倉, 大山, 林: 日本語語彙大系, 岩波書店 (1997)
- [5] 長尾真 (編): 自然言語処理, 岩波ソフトウェア科学, No.15(1996).