自然言語処理 133－2
（1999．9．10）

非対訳コーパスを用いた日本語複合名詞の英訳語推定

イラム シャハザド † 大竹 清敬 † 増山 繁 † 山本 和英 ‡

{iram, otake, masuyama}@smlab.tutkie.tut.ac.jp yamamoto@itl.atr.co.jp

† 豊橋技術科学大学 知識情報工学系
‡ ATR 音声翻訳通信研究所

概要

日本語では，複数の名詞を組み合わせることにより複合名詞が限りなく作り出される．従って，このような複合名詞を人手で辞書に収録することには限界がある．そこで，機械翻訳システムにおいては，これらの複合名詞の対訳を自動的に獲得する必要がある．複合名詞の構成語の訳語を利用する手法では，日英で構成語の対応がない複合名詞に対しては対処できない．本手法では，複合名詞の文脈を利用して，英語のコーパスから対訳が出現する可能性のある部分を特定し，ヒューリスティックス等を用いてその部分から対訳を抽出する．

Identifying English Equivalents of Japanese Compound Nouns Using Non-aligned Corpora

Iram SHAHZAD,† Kiyonori OHTAKE,† Shigeru MASUYAMA,†
and Kazuhide YAMAMOTO‡

† Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.
‡ ATR Interpreting Telecommunications Research Laboratories

Abstract

Japanese documents contain a lot of compound nouns. These compound nouns are created from different combinations of nouns and are too large in number to be contained in a manually-created dictionary. Automatic extraction of their translations from English corpus is highly desirable. Utilizing the translations of the constituent parts is not effective for those compound nouns which do not have correspondence with their translations in part-by-part basis. This leads to the necessity of a different approach for these kinds of compound nouns. Our method, at first, detects the parts of English corpus which may contain the translation by using the context of the compound noun, and then, identifies the translation using some heuristics.

1 はじめに

国際化，情報化の進展に伴い，機械翻訳の重要性が高まってきている．しかし，自然言語は多彩で柔軟な表現構造を持つので，翻訳は依然として難しい．特に，日本語から英語への翻訳の場合は様々な言語現象が難しさの原因となる．複合名詞はその一つである．

日本語は極めて造語力の強い言語であるため，複数の単語からなる複合語が頻繁に使用される．これは，既存の語を組み合わせて新しい概念を効率よく表現する方法である．複合語の中に大きな割合を占めているのは複合名詞である．このような名詞は，文の内容を把握するための重要な語でもあり，機械翻訳の際，正確に訳される必要がある．しかし，生成される複合名詞の数には限りがないため，全ての複合名詞を辞書に収録することはできない．従って，機械翻訳システムにおいては，日本語複合名詞の対訳を自動的に獲得する必要が生じる．

笹岡らは，複合名詞の対訳生成に，要素合成の原理を用いている [1]．しかし，このような手法は極めて限定された複合名詞に対してしか適用できない．田中らは，複合名詞の構成語の訳語を手掛かりに，非対訳コーパスから複合名詞の対訳を獲得している [2]．しかし，この方法では，構成語の対応がない複合名詞に対しては対処ができない．そこで，このような複合名詞に対して，構成語の訳語を利用せずに複合名詞を一つの未知語と見なし，その対訳の獲得を行うことが必要である．

対訳コーパスから未知語の対訳を獲得する研究として，[3, 4, 5, 6, 7, 8, 9, 10] がある．しかし，対訳コーパスの作成には非常にコストがかかる．

非対訳のコーパスを使って未知語の対訳を獲得する試みもある [11, 12, 13, 14, 15, 16]．これらの研究では，原言語と目的言語間での文脈 (共起語で近似的に表される) の類似性を使って，未知語の対訳を推定しているが，目的言語のコーパス上のあらゆる単語の共起語との類似性を求めているため，効率などの問題点が残っている．

本研究で対象とするのは，既存の辞書に記載されておらずしかも日英では構成語の対応がないような複合名詞である．また，英訳語が 1 語である複合名詞もあるが，そのような語は英和辞書に記載されている可能性が高いため，対象外とする．本研究でも，文脈の類似性という概念を対訳推定に利用する．つまり，本手法は，「日本語のコーパスにおいて，ある複合名詞が出現する文脈と，英語のコーパスにおいて，その訳語が出現する文脈との間は共通性がある」という仮定に基づく．文脈を特定する要素として今回は，「どのような語と共起するか」に加えて「どのような語とどのような関係を持つか」を利用する．また，従来の手法とは異なり，本研究では，最初に対訳の検索範囲を限定する工夫をする．限定された範囲内で，対訳を検索することによって効率の向上を目指す．

2 複合名詞対訳推定手法

日本語のコーパスにおいて，ある複合名詞が出現する文脈と，英語のコーパスにおいて，その訳語が出現する文脈との間には類似性があると考えられる．本手法では，この類似性を利用して，複合名詞の対訳を推定する．ある語の「文脈」を，その語が

- 他のどのような語と共起し，
- 他のどのような語とどのような接続関係を持つか

で近似する．ここで，「共起する」とは，語対が一定の距離以内に一定以上の確率で出現することと定義する．また，語との「接続関係」を，その語とどのような機能語を介して接続されているかで表す．

本稿の対訳推定手法は，

1. 対訳を求めようとする複合名詞の文脈を特定する「文脈特定部」，
2. 文脈情報を同等な英語に変換する「文脈変換部」，

3. 英語コーパス上で複合名詞の対訳が出現する可能性が高い領域 (以下，対訳検索領域と呼ぶ) を特定する「対訳検索領域特定部」，

4. 対訳検索領域から対訳を特定する「対訳特定部」

から成る．

2.1 文脈特定部

ここで，複合名詞に対して，以下で述べる「共起語集合」と「接続関係集合」を獲得する．

2.1.1 共起語集合の獲得

複合名詞の周辺に出現する語の中には，複合名詞と強い関係を持って出現するものとそうでないものがある．共起関係の強度の尺度として，様々なものが利用されている [9, 10, 17]．本手法では，周辺に出現する語の中から関係の弱いものを排除するためにある複合名詞 A に対する別の語 B(名詞，動詞，形容詞に限定) の共起強度 $CS_A(B)$ を次式で計算し，$CS_A(B) \geq 50\%$ である B を共起語とする．

$$\text{共起強度 } CS_A(B) = \frac{f(A,B)}{f(B)} \times 100\ (\%)$$

ただし，$f(B)$ はコーパス全体における B の出現回数，$f(A,B)$ は A が含まれる文とその前後 1 文の範囲に B が出現する回数を表す．

対訳を求めようとする複合名詞の共起語集合の獲得は，コーパス全体にわたって得られた共起語をまとめることによって行う．

2.1.2 接続関係集合の獲得

対訳を求めようとする複合名詞と，それが含まれる文脈内の全ての語との関係を獲得することは困難であり，さらにそれらを別の言語に対応させることも困難なので，ここでは，単純で表層的な関係を用いる．

複合名詞 A が他の語 C(名詞，動詞，形容詞に限定) と助詞 P を介して，接続されているとき，接続関係を 3 項によって定義し，$\langle C,\ P,\ S\rangle$ と表記する．ここで S は，正負の符号を意味する．A と C の接続パターンは APC と CPA の 2 通り考えられるので，それらを区別す[1]るために S を導入した．表 1 で示すような形で接続関係を獲得する．助詞として，「が」，「は」，「を」，「に」，「で」，「から」，「と」，「へ」，「の」と ϕ を考慮する．ϕ は長さ 0 の文字列を意味する．このようにして得られた接続関係をまとめることにより接続関係集合を得る．

表 1: 接続関係

接続パターン	接続関係
APC	$\langle C,\ P,\ +\rangle$
CPA	$\langle C,\ P,\ -\rangle$

2.2 文脈変換部

ここで，英語コーパス上から複合名詞の対訳を推定する準備として，共起語集合と接続関係集合の要素を同等な英語 (それぞれ，「手掛かり語集合」と「手掛かり関係集合」と呼ぶ) に変換する．

2.2.1 手掛かり語集合の獲得

共起語集合の要素の中で既存の和英辞書から訳語が獲得できるものは，訳語を求める[1]．これらの訳語を要素として手掛かり語集合をつくる．

2.2.2 手掛かり関係集合の獲得

接続関係を表 2 に示すテーブルを用いて，英語の同等なものに変換する．ただし，テーブル中の C' は，既存の辞書から得られる C の訳語[1]，$\bar{S}$ は，S を反転したものを表す．このように変換された接続関係を要素として手掛かり関係集合をつくる．

[1]ひとつの語に対する訳語が複数ある場合は，本来ならばその中から適切なものを決定する必要がある．しかし，ここでは全ての訳語を採用する．

表 2: 接続関係変換テーブル

日本語	英語
$\langle C$, が, $S\rangle$	$\langle C'$, $(is\|are\|was\|were)$, $S\rangle$
$\langle C$, は, $S\rangle$	$\langle C'$, $(is\|are\|was\|were)$, $S\rangle$
$\langle C$, を, $S\rangle$	$\langle C'$, ϕ, $\bar{S}\rangle$
$\langle C$, に, $S\rangle$	$\langle C'$, $(to\|in)$, $\bar{S}\rangle$
$\langle C$, で, $S\rangle$	$\langle C'$, $(by\|with)$, $\bar{S}\rangle$
$\langle C$, から, $S\rangle$	$\langle C'$, $from$, $\bar{S}\rangle$
$\langle C$, と, $S\rangle$	$\langle C'$, and, $S\rangle$
$\langle C$, と, $S\rangle$	$\langle C'$, $with$, $\bar{S}\rangle$
$\langle C$, へ, $S\rangle$	$\langle C'$, $(to\|at)$, $\bar{S}\rangle$
$\langle C$, へ, $S\rangle$	$\langle C'$, at, $\bar{S}\rangle$
$\langle C$, の, $S\rangle$	$\langle C'$, ϕ, $S\rangle$
$\langle C$, の, $S\rangle$	$\langle C'$, of, $\bar{S}\rangle$
$\langle C$, ϕ, $S\rangle$	$\langle C'$, ϕ, $S\rangle$
$\langle C$, ϕ, $S\rangle$	$\langle C'$, of, $\bar{S}\rangle$

2.3 対訳検索領域特定部

英語のコーパスで，3 文を単位とした領域ごとに，手掛かり語集合との重なりの度合を求める．これは，領域内に手掛かり語集合の要素が出現する回数である．この値は 3 以上である全ての領域を対訳検索領域として採用する．

2.4 対訳特定部

得られた領域群の文を要素として，文集合とする．文集合の文から以下の全ての条件を満たす部分を対訳候補として採用する．

条件：

1. 先頭語は，名詞か形容詞である．
2. 末尾語は，名詞である．
3. 先頭語と末尾語の間には，名詞，形容詞，前置詞，限定詞，“and” 以外は存在しない．
4. 構成語数は，2 以上である．(日本語複合名詞の英訳語が 1 語である場合もあり得るが，そのような語が英和辞書に記載されている可能性が高いため，ここで対象外としている．)
5. 構成語数は，(入力複合名詞の構成語の数×3) 以下である．

このようにして得られた対訳候補から，次のような処理で，入力された複合名詞の対訳である可能性が低いものを候補から排除する．

1. 対訳候補の構成語[2]の日本語訳語[3]を組み合わせて作成した文字列を収集する．
2. 収集された文字列群の中に，対訳を求めようとする複合名詞と一致せずにかつ日本語コーパス上に存在するようなものがある場合は，その対訳候補を削除する．

残された各対訳候補に対して，以下の 2 つの方法で重み付けをする．

1. 文集合中において，対訳候補が出現する回数を重み W_1 とする．
2. 対訳候補が，手掛かり関係集合中の関係を英語のコーパス上で再現する回数を重み W_2 とする．

次に，重み $W = W_1 + W_2$ を計算する．最後に，最高の W を持つ候補を複合名詞の対訳として採用する．

3 実験と考察

本手法の検証のために実験を行った．実験では，本手法による該当文 (正解を含む文) の特定率，最終段階での対訳候補数 (X) とその中での正解の順位 (Y) を求め，手法の評価を行った．該当文の特定率を以下に定義する適合率と再現率で求めた．

$$\text{適合率 } P = \frac{\text{特定された該当文数}}{\text{特定された文数}} \times 100\ (\%)$$

$$\text{再現率 } R = \frac{\text{特定された該当文数}}{\text{全コーパス中の該当文数}} \times 100\ (\%)$$

[2]名詞，動詞，形容詞だけを考慮する．
[3]末尾の「する」,「な」,「の」を削除したもの

ここで，文が特定されるとは，2.4節の文集合中に含まれていることである．本実験では，日本語のコーパスとして，日本経済新聞1996年の全記事(約184万文，約168MB)を使用した．英語のコーパスとして，WWWから取得した「経済」に関する記事群(約54万文，約40MB)を使用した．辞書として，EDR日英対訳辞書(見出し数：約36万)，EDR英日対訳辞書(見出し数：約29万) [18] およびEdict(日本語見出し数：約7万，英語見出し数：約11万) [19] を利用した．日本語の形態素解析にJUMAN [20]，英語の品詞獲得にBrill Tagger [21] を使用した．

今回は，コーパスの量が不十分なため，コーパスに比較的多く含まれる複合名詞(経済・金融に関連のある10個の用語)を対象として実験を行った．また，既存の辞書での記載の有無または日英での構成語間の対応の有無は本手法の結果に影響を与えないので，入力データの選択にそのような条件を考慮しなかった．実験結果を表3に示す．また，表4に「外資系企業」に対して推定された英訳語候補を示す．

表 3: 実験結果

複合名詞	P	R	X	Y
経済政策	38	71	19	3
経済協力	21	37	88	25
金融機関	45	60	15	5
世界貿易	41	72	9	4
為替相場	28	48	27	11
公定歩合	38	51	16	5
先物取引	47	63	13	4
景気予測	24	66	21	7
外資系企業	36	78	6	2
記者会見	23	56	11	7

再現率の平均値は60%，適合率の平均値は34%である．本研究の目的から見れば，再現率はあまり重要ではなく，適合率の方が重要である．今回の実験では，適合率は低いため，対訳候補に不正解を多く含んでしまう．しかし，多数の対訳候補の中で正解が上位に存在しているのは，対訳特定部の有効性を示している．

表 4: 対訳抽出例

外資系企業
Managing Director
foreign company
debt restructuring
Miguel board
Seamico Securities
general manager

4 おわりに

本研究では，日本語と英語で構成語の対応がない日本語複合名詞の対訳推定を目的として，非対訳のコーパスを用いて文脈情報を手掛かりにそのような対訳の推定を試みた．本手法では，まず，対訳が出現する可能性のある領域を特定し，次に，特定された領域から対訳を推定した．この手法は，対訳候補を絞るのにある程度有効だという結果が得られた．しかし，対訳候補の中に多数の不正解も含まれてしまうため，今後改善が必要である．今後の課題としては以下の点があげられる．

1. さらに質の良い共起語を獲得する方法の検討．
2. 語の文脈情報をさらに有効に利用．
3. 手掛かり語集合を求める際の多義性への対処．
4. 英語のコーパス上でマッチングを行う際，単語の原形を求める処理などを加えることによるマッチング率の向上．
5. より大きい英語のコーパスを使っての実験．
6. さらなる評価実験．

参考文献

[1] 笹岡久行, 荒木健次, 桃内佳雄, 栃内香次：帰納的学習を用いた訳語推定手法における単語片対の抽出元の選択数に関する性能評価, 言語処理学会 第5回年次大会 発表論文集, pp. 357–360 (1999).

[2] 田中貴秋, 松尾義博：対訳関係のないコーパスからの複合名詞対訳の獲得, 言語処理学会 第5回年次大会 発表論文集, pp. 29–32 (1999).

[3] 大森久美子, 佐藤健吾, 中西正和：共起関係を利用した対訳コーパスからの連語の対訳表現抽出, 情報処理学会研究報告 NL-122, pp. 13–20 (1997).

[4] 佐藤健吾, 中西正和：最大エントロピー法による対訳単語対の抽出, 情報処理学会研究報告 NL-122, pp. 21–27 (1997).

[5] 米沢恵司, 松本裕治：漸進的対応付けによる対訳テキストからの翻訳表現の抽出, 言語処理学会 第4回年次大会 発表論文集, pp. 576–579 (1998).

[6] Kupiec, J.: An Algorithm for Finding Noun Phrase Correspondences in Billingual Corpora, in *Proceedings of ACL-93*, pp. 17–22 (1993).

[7] Kaji, H. and Aizono, T.: Extracting Word Correspondences from Bilingual Corpora Based on Word Co-occurrence Information, in *Proceedings of COLING-96*, pp. 23–28 (1996).

[8] Kumano, A. and Hirakawa, H.: Building an MT dictionary from parallel texts based on linguistics and statistical information, in *Proceedings of COLING-94*, pp. 76–81 (1994).

[9] 北村美穂子, 松本裕治：対訳コーパス中の共起頻度に基づく対訳表現の自動抽出, 情報処理学会研究報告 NL-114, pp. 69–76 (1996).

[10] Smadja, F., Mckeown, K. R. and Hatzivassiloglou, V.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, *Computational Linguistics*, Vol. 22, No. 1, pp. 1–38 (1996).

[11] 辻慶太, 芳鐘冬樹, 影浦峡：対訳コーパスからの訳語対抽出における辞書情報の利用について, 言語処理学会 第5回年次大会 発表論文集, pp. 402–405 (1999).

[12] 麻野間直樹, 中岩浩巳：目的言語の単語共起情報を利用した訳語選択と未知語の訳出, 言語処理学会 第5回年次大会 発表論文集, pp. 442–445 (1999).

[13] Tanaka, K. and Iwasaki, H.: Extraction of Lexical Translations from Non-aligned Corpora, in *Proceedings of COLING-96*, pp. 580–585 (1996).

[14] Fung, P. and Yee, L. Y.: An IR Approach for Translating New Words from Nonparallel, Comparable Texts, in *Proceedings of COLING-ACL '98*, pp. 414–420 (1998).

[15] Rapp, R.: Automatic Identification of Word Translations from Unrelated English and German Corpora, in *Proceedings of ACL-99*, pp. 519–526 (1999).

[16] Kikui, G.: Term-list Translation using Monolingual Word Co-occurrence, in *Proceedings of COLING-ACL '98*, pp. 670–674 (1998).

[17] Niwa, Y. and Nitta, Y.: Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries, in *Proceedings of COLING-94*, pp. 304–309 (1994).

[18] (株) 日本電子化辞書研究所：EDR 電子化辞書 1.5 版 使用説明書 (1996).

[19] Breen, J. W.: *EDICT, Freeware Japanese/English Dictionary, V99-002* (1999).

[20] 黒橋禎夫, 長尾真：日本語形態素解析システム JUMAN version 3.6 (1998).

[21] Brill, E.: *Rule Based Part of Speech Tagger, Version 1.14* (1994).