対訳関係のないコーパスからの複合名詞対訳の獲得

田中 貴秋　　松尾 義博
NTTコミュニケーション科学基礎研究所
{takaaki,yosihiro}@cslab.kecl.ntt.co.jp

1 はじめに

自然言語処理を行う際に，扱いが困難な問題として複合語がある．ある複数の単語から構成されている語句を，一語の複合語と扱うか，独立した語の並びとして扱うかの区別は曖昧であるが，機械翻訳を目的と考えた場合には慣用的なものも含めて繰り返し使用される語の組合せからなる表現は辞書として持っておく方が都合が良い．たとえば，「個人投資家」という表現を英語に翻訳する場合には，「個人」と「投資家」と分割して考えると「個人」に対して “person” , “personal”, “individual” のどれが適切であるのか，あるいは他の語が適切であるのかを判断することが難しく，「個人投資家」と “individual investor” の対訳を辞書に持つ方が扱いやすい．

Fungは，文単位などでの対応付けのなされていない対訳コーパスからパターンマッチングの手法を使って名詞，固有名詞の対訳表現を収集している [1]．この方法では辞書で直接対応付けを行わないので，収集される表現は辞書の制約を受けないが，両言語コーパスでの位置情報を用いるため対訳関係のないコーパスには適用できない．また，Smadjaらは，対訳コーパスから統計的方法を用いて2言語の collocation を獲得する方法を提案している [2]．米沢らは，対訳コーパスから単語列を抽出し2言語間の共起関係を使用して単語列の対応付けを行い翻訳辞書を構成する方法を提案している [3]．これらの方法も既存の辞書による制約はないが，文対応の付けられた対訳コーパスを用いることが前提となっている．

しかし，直接対訳関係のあるコーパス間でなくても，内容が同一の分野に関連するコーパスであれば，専門用語など共通に使われる表現が双方のコーパスに含まれている可能性が高いと考えられる．各言語コーパスから複合語の候補を抽出し，それらを言語間で対応付けることができれば，新たな対訳表現を獲得できる．

我々は対訳関係のないコーパスから正解の対訳例に類似した対訳表現を収集する方法を提案した [4]．この方法では各コーパスに含まれる対訳例に類似した表現を対訳辞書で対応づけ，統計情報を使って候補の絞り込みを行う．また，Daganらは，ある言語の依存関係にある語の組合せ[1] (*subject-verb*,*verb-object* など) に対応する他の言語の組合せを，目的言語のコーパスのみを使って求めている [5]．原言語の表現と目的言語の表現が構成語のレベルで対応する場合には，構成語の訳語選択 (あるいは語義の曖昧性解消) の問題ととらえてこの方法を複合語の対訳を収集に適用することが考えられる．

しかし，これらの方法では構成語間に通常辞書に記載されているような対訳関係がない表現は収集できない．例えば，「経常利益」と “current profits” のような対訳は，通常，「経常」と “current” という語が対訳関係にないために獲得できない．

本稿では，対訳関係のない異なる言語コーパスから対訳となる複合名詞表現を獲得する方法を提案する．対訳辞書，シソーラスを使って，対訳表現を収集する方法とその実験結果について述べる．

2 専門用語辞書

2.1 収録語の特徴

コーパスに含まれている専門用語などの複合名詞は内容の分野によって大きく変化する．そのため機械翻訳システムを特定の分野に適用する際には，専門用語の対訳辞書を持つことが必要となる．しかし，分野ごとに用語を収集し辞書を構築することは膨大な作業量を必要とするため，専門用語の特徴を利用して機械的に収集することが期待される．

既存の専門用語辞書に記載されている語の特徴を見るために構成する単語の品詞並びを調べた．インタープレス版ビジネス用語辞書 (見出し語約106,000語，以下単に専門用語辞書と記す) の見出し語 (日本語) と訳語 (英語) をそれぞれALTJAWS[2]，Brill Tagger[7] を使って形態素解析を行い，結果を品詞並びによって集計したものが表1,2である．表では，品詞は，名詞を **N**，接辞 (日本語) を **X**，形容詞を **J**，前置詞 (英語) を **P** で表している．日本語では2語以上から構成される表現の約43% が **NN** の組み合わせであることがわかる．また，英語では名詞2語以上からなる表現のうちで **NN**，**JN** の組み合わせがほぼ同程度で両者を合わせて53% を占めている．本稿では，これらの多く現れる品詞列を各言語のコーパスから抽出し両者を対応付けることを考える．

2.2 収録語とコーパス

特定の分野の内容について書かれた文章であれば，書かれた言語が異なっていても共通の概念を表す専門用語が使われていると考えられる．

表3は，専門用語辞書中の日英対訳のうちで，日本語と英語がともに名詞2語の連続 **NN** で構成されている表現について，日本経済新聞1年分[3]，と Wall Street Journal 1年分に現れたものの数を示している．収録されている **NN** 表現のうちどちらかがコーパスに出現し

[1] syntactic tuples と呼んでいる．

[2] 機械翻訳システム ALT-J/E[6] の形態素解析器

[3] 日本経済新聞 CD-ROM 94版を使用した

表 1: 辞書に収録されている専門用語 (日本語)

N	34644	(32.8%)
NN	30912	(29.3%)
NX	9903	(9.4%)
NNN	6714	(6.3%)
NNX	4667	(4.4%)
その他	18770	(17.8%)
計	105610	(100%)

表 2: 辞書に収録されている専門用語 (英語)

N	33040	(31.3%)
NN	20285	(19.2%)
JN	18316	(17.3%)
NPN	4202	(4.0%)
NNN	2853	(2.7%)
JNN	2816	(2.7%)
その他	24098	(22.8%)
計	105610	(100%)

たものは約1/4にのぼっている．さらにこのうち2000組弱がコーパス中に出現している．この種類のものは単語レベルで2言語の対応をつけることができれば，コーパス中から獲得することができると期待される．

表 3: 辞書に収録されている専門用語のコーパス中出現数

日本語表現	8126
英語表現	4357
日英ともに出現	1962 組

3 複合名詞の単語レベル対応

辞書中に含まれている複合名詞を，日本語と英語の単語の対応関係で分類すると以下のような種類がある．

1. 両言語が同種の品詞から構成され，単語辞書により一対一に対応がとれるもの
   (例) 証券 /N 会社 /N : securities/N company/N
2. 一部の品詞が異なるが，派生語に直せば一対一に対応のとれるもの
   (例) 技術 /N 革新 /N : technical/J innovation/N
3. 付属語などを含むが，内容語の対応が一対一にとれるもの
   (例) 生産 /N 能力 /N : capacity/N of/P production/N
4. 一部の内容語の対応がとれるもの
   (例) 営業 /N 利益 /N : operating/J profit/N
5. 対応のとれないもの
   (例) 違憲 /N 立法 /N 審査 /N 権 /X : judicial/J review/N

1.〜3. についてはコーパスから対象とする単語列を抽出して二言語間で辞書対応のとれるものを選び出せば，対訳表現の候補を収集できると考えられる．例えば，「生産」の英訳候補に (product, production, output) が，「調整」には (adjustment, correction) があった場合，単語列の中に「生産調整」と "production adjustment" が存在すれば単語間の対応がとれるので，この対を対訳表現の候補とすることができる．

4. のように一部の語のみしか対応付けらない表現についても残りの語に何らかの対応関係を見い出すことができれば，収集可能であると考えられる．辞書で対応のとれなかった残りの語の関係には次のようなものが含まれる．

(i) 本来対訳として扱うべきであるが，辞書収録語の不足により対応がとれなかったもの

(ii) 類似の意味をもつもの
(例) 経営 計画 - business plan

(iii) 複合語となったときに現れる訳語
(例) 経常 利益 - current profits

ある語句に対する訳語を目的言語コーパスを使用して求める方法は，訳語候補となるものが辞書に存在していることが前提となっているが，現実の辞書には見出し語に対して訳語となるべき語が必ずしも網羅されていない．特に機械翻訳システム用の辞書は，通常訳語選択の困難さからやみくもに対訳語を登録することは行われない．しかし本稿の目的ような場合には，ある語句に対して対訳になりうる可能性のある語句はある程度広い範囲で収録されていることが望ましい．

(i)のように複数の辞書を使用することにより補うことができるものもあるが，(ii)，(iii)は，通常の辞書には単語の対訳としては記載されていない単語対応が含まれる．そのうち (ii). は類義語辞書などにより対応する語の範囲を広げることにより対応付けが可能なものである．(ii)の例では「経営」と "business" の訳語である「営業」の類義性が利用できる．類義語による対応によって対応付けられる語の範囲が広くなりすぎる恐れがあるが，複合語の場合は構成する語全体の対応関係から判断することで絞り込みを行えると考えられる．

また，(iii)のようにこのような類義性は利用できないものも存在する．(iii)の例では，「経常」と "current" の間では直接意味的な対応関係はない．しかし，日本語の「経常」を含んだ複合名詞の英語訳には以下のように "current" が良く使われる．このように複合名詞として使用されたときに対応する可能性のある語は対訳語を検索する手がかりとすることができる．

経常収支	**current** account
経常取引	**current** transactions
経常価格	**current** price
経常費	**current** expenses

4 日英対訳語の収集方法

2.1で述べたように，それぞれの言語で専門用語辞書に収録されている用語の品詞列のうち頻度の高いものは種類が限られている．コーパス中からこの品詞列から成る単語列を抜き出すことにより，複合名詞を含んだ表現を収集することができる．単純に抜き出された単語列には複合名詞として不適切なものも多く含まれるが，これらの表現の多くは3で述べたような対応付けが行えず除去できると考えられる．構成する単語が対応付けられたものを選び出すことにより，対訳表現の候補となる表現が獲得される．

本手法は以下の過程からなる．

(1) 各コーパスから収集対象とする品詞並びからなる単語列を抽出する

(2) 抽出された単語列を言語間で対応付けを行う

(3) 対応付けの結果から確からしいものを選出する

4.1 単語列の収集

コーパスは形態素解析器により分かち書きをして品詞を付与したものを使用する．処理を簡単にするため単語列の収集は単純な品詞列情報のみによって行う．

4.2 言語間の対応付け

本手法では，二言語の単語の対応として以下の (a) 〜 (c) の3種類を考える．

(a) 対訳辞書による対応 和英辞書などに見出し語と訳語として登録されている単語の関係に対して与える対応付けである．日本語の名詞が英語では形容詞に変わる場合があるため他品詞の派生語を含めて対応をとる．

(b) 類義性による対応 本稿では，類似性の尺度に日本語語彙大系 [8] の意味カテゴリを用いた．語彙大系では各名詞に対して木構造からなる約3000の意味属性が付与されている．英語の属性は対訳辞書で対応する日本語の意味属性を使用した．日本語の単語 w_J と英語の単語 w_E に同じ意味カテゴリが与えられているときに対応させる．各名詞には複数の意味属性が付与されており，さらに英語には複数の名詞が対応するので，いずれかのカテゴリが一致した場合に対応付けを行った．木構造を利用して類似度を計算することも考えられるが，カテゴリの組合せが複数存在し，対応付ける範囲を広げ過ぎるので本稿では同一カテゴリに限った．

(c) 複合語対応辞書による対応 専門用語辞書の対訳から複合語対応辞書（以下対応辞書と記す）を作成し，新たな対応付けを行う．辞書中の複合名詞を構成する単語間を対訳辞書により対応付けを行い，対応付けが行われずに残った単語の組を集め対応辞書に登録する．例えば，「経常利益」と“current profits”という対訳について (利益,profits) のみ単語対応が付いた場合，残りの組 (経常,current) を対応辞書に加える．

以上の単語対応に単語列間の対応尤度を決定する．日本語の単語列 J_x=$(w_{J1},...,w_{Jm})$ と，英語の単語列 E_y=$(w_{E2},...,w_{En})$ との間の対応尤度 $cor(J_x,E_y)$ を (1) 式のように定義する．単語間の対応は対訳辞書対応が最も高くなるよう対応度 $link()$ を設定する．また，主名詞と修飾語間の対応を低くするために，各対応度には重み $wg()$ を乗じている ((3) 式)．日本語の名詞列 (w_{J1},w_{J2}) と英語の名詞列 (w_{E1},w_{E2}) の場合には w_{J2} と w_{E2} が主名詞であることが多いので，w_{J1} と w_{E2}，w_{J2} と w_{E1} の対応は低くする．

$$cor(J_x,E_y)=\frac{links_{JE}+links_{EJ}}{n(J_x)+n(E_y)} \tag{1}$$

ここで，

$n()$: 単語列に含まれる単語数．

$links_{JE}=\sum_i max_j(link(w_{Ji},w_{Ej})wg(w_{Ji},w_{Ej}))$

$links_{EJ}=\sum_j max_i(link(w_{Ji},w_{Ej})wg(w_{Ji},w_{Ej}))$

$$link(w_a,w_b)=\begin{cases}1 & :\text{対訳辞書対応}\\ \lambda_s(0<\lambda_s<1) & :\text{類義性対応}\\ \lambda_d(0<\lambda_d<1) & :\text{対応辞書対応}\\ 0 & :\text{対応なし}\end{cases} \tag{2}$$

$$wg(w_a,w_b)=\begin{cases}1 & :\ w_a,w_b\ \text{がともに主名詞 or}\\ & \ \ w_a,w_b\ \text{がともに修飾語}\\ \alpha(0<\alpha<1) & :\text{その他}\end{cases} \tag{3}$$

5 実験と考察

コーパスとして日本経済新聞 (約203万文, 約200MB), Wall Street Journal(約250万文, 約150MB) を使用して実験を行った．

日本語と英語がともにNNで構成される表現を収集対象とした．使用するコーパスに含まれるNN表現は，日本語が882,250種類，英語が419,928種類であった．専門用語辞書中の日本語NNと英語NNである対訳表現のうちコーパスに出現した1962組を正解の対訳表現 (以後正解対訳セットと記す) として用いた．本手法の対応づけ方法を評価するためにこの正解セットの日本語1693種を選び，英語のNNの表現約42万種類に対して4.2で述べた (a) 対訳辞書の対応付けを行った (α=0.7, で対応尤度0.9以上)．結果を表4に示す．再現率は 32.6% と低いが，これは既存の対訳辞書のみの対応付けでは不十分であることを示している．また，不正解となった対訳の中には一つの日本語に対して正解以外の複数の英語候補が選ばれたものが含まれる．これらは表5のように表現の出現頻度により順位付けをすることが考えられる．

次に他の二つの対応付け方法を組み合わせた場合について実験した．正解セットの中から日本語の出現頻度の高い221組を選び，(a) 対訳辞書単独の対応付け，(a) 対訳辞書 +(b) 類義性による対応付け，(a) 対訳辞書 +(c) 対応辞書による対応付け，それぞれで対訳を収集した．(b) には日本語語彙大系の意味カテゴリを用い，(c) の複合語対応辞書による対応付けには，専門用語辞書中収録の日本語と英語がともに2語からなっている対訳から正解対訳セットを除き，残りの表現を対訳辞書によって対応付けを行って，対応のとれなかった

表 4: (a) 対訳辞書対応で収集した結果

	(a)
収集された表現対	1,063
うち正解	641
再現率	32.6%
適合率	60.0%

表 5: 一つの日本語に複数の英語候補が選ばれた例

日本語	英語	出現頻度	辞書記載
設備投資	equipment investment	13	○
	facility investment	1	
為替差益	deficit reduction	60	○
	deficit cut	6	
	loss cut	2	

単語の組を集めて新たに対応辞書を作成した．(2) 式の $\lambda_s = 0.9$，$\lambda_d = 0.9$ とし対応尤度が 0.9 以上のものを対訳として収集した．収集された対訳候補の例を表 6 に示す．

表 7 に示すように，(a)+(b) では再現率が最も高く広く収集されていることがわかる．反面，候補の収集が多くなり適合率を下げる結果となった．これは使用したシソーラスが日本語側のものであり，英語の意味カテゴリは日本語の訳語から参照しているために対応するカテゴリが広がり過ぎたことが一因であると考えられる．ただし，いずれの実験においても単数形と複数形を別の一語と数えているが辞書には単複通常どちらかしか載せていないことや，辞書と全く同一の表現ではないが対訳として認められるようなものも含まれているので実質の正解はこの結果より高い．

(a)+(c) では再現率，適合率ともに (a) 単独，(a)+(c) の中間の値を示しており対応辞書は，類義性対応に比較して妥当な対応付けを行っているといえる．これは，意味カテゴリが一般的な広い対応を与えるのに対し，対応辞書は一分野に現れる対応関係を与えるためと考えられる．(a)+(c) については，正解セット全体についても実験を行った (表 8)．正解対訳セットの 60% 以上が収集され，再現率は (a) 対訳辞書単独の場合の約 2 倍になっている．既存の辞書から対応辞書を作成することによりより多くの対訳表現が収集可能になることが確かめられた．

6 おわりに

対訳関係のない 2 言語のコーパスから抽出した単語列を辞書や意味属性を用いて対応付けを行い複合名詞対訳を収集する方法を提案した．対訳辞書と，単語の類義性，複合名詞の構成単語の対応関係を利用することにより広範囲に対訳を獲得できることを示した．対応付け候補を広げる方向で進めたため適合率が下がる傾向があったが，3 種類の対応付けの組合せや出現頻度などの統計情報を用いて絞り込みを行うことが課題である．

表 6: 収集された対訳候補

日本語	英語	出現頻度	対応種類	辞書記載
株式売買	stock market	3598	(a)(b)	○
	stock trading	115	(a)(b)	
	stock trade	2	(a)(a)	
基本計画	master plan	11	(c)(a)	○
終身雇用	life employment	1	(c)(a)	○
転換価格	conversion value	2	(c)(a)	○
	conversion price	36	(c)(a)	
輸入採算	import cost	4	(a)(c)	○

表 7: 対応付け方法による比較

	(a) 単独	(a)+(b)	(a)+(c)
収集された表現対	152	1,437	705
うち正解	79	189	136
再現率	35.7%	85.5%	61.5%
適合率	52.0%	13.2%	19.3%

表 8: (c) 複合語対応辞書対応で収集した結果

	(a)+(c)
収集された表現対	4,847
うち正解	1,176
再現率	60.0%
適合率	24.2%

参考文献

[1] Pascale Fung. A pattern method for finding noun and proper noun translation from noisy parallel corpora. In *Proc. of the 33rd Annual Conference of Association for Computational Linguistics*, pp. 236–243, 1995.

[2] Frank Smadja, Kathleen R. McKeown, Vasileios Hatzivassiloglou. Translation collocations for bilingual lexicons: A statistical approach. *Journal of Association for Computational Linguistics*, pp. 1–38, 1996.

[3] 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, pp. 576–579, 1998.

[4] 田中貴秋, 松尾義博, 大山芳史. 非対訳コーパスからの日英機械翻訳ルールの自動獲得. 言語処理学会第 4 回年次大会, pp. 594–597, 1998.

[5] Ido Dagan, Alon Itai. Word sense disambiguation using a second language monolingual corpus. *Journal of Association for Computational Linguistics*, pp. 563–596, 1994.

[6] Satoru Ikehara. Multi-level machine translation method. *Journal of Future Computing Systems*, Vol. 2, No. 3, 1989.

[7] Eric Brill. A corpus-based approach to language learning. 1993.

[8] 池原悟, 宮崎正弘, 白井 諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙大系. 岩波書店, 1997.