日英複合名詞の対応判別

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

様々な言語で書かれた膨大な量の電子テキストがネットワークなど様々な媒体を通して流通するようになり，機械翻訳やクロスリンガル情報検索など言語を横断した自然言語処理への関心が高まっている．異なる複数の言語を扱う言語処理で最も基本的で重要な情報が対訳辞書である．主要な言語の基本となる語彙は既存の対訳辞書である程度の部分がカバーされていると考えられるが，新語や専門用語などは必要に応じて新たに収集しなければならない．しかし，それらを集めることは容易ではなく，特に複合語はあらゆる単語の組み合わせから無限に作られるために網羅することが困難である．

そこで，対訳となる表現をコーパスから機械的に収集する方法が研究されている．対訳コーパスを使用する方法と，対訳関係を持たないコーパスを組み合わせて用いる方法 [1, 2] の2つに大別される．前者は記述されている内容の対応がとりやすいコーパスであれば高い精度が期待できるが，使用できるコーパスがかなり限定されてしまう．それに対し後者の方法では，対訳コーパスを使用した場合ほどの精度を達成するのは難しいが，あらゆるコーパスを組み合わせて利用できるという利点がある．

筆者らは，対訳関係のない日英のコーパスから意味的に対応する語を獲得する研究を行っている [3]．この方法では，与えられた日本語の複合名詞に対する英語訳候補を構成語の情報をもとにコーパスから抽出し，もっともらしい候補を選出する．本稿では，複合名詞の対訳を獲得することを目的として各コーパスから収集した文脈情報（共起情報）を用いてもとの日本語と意味的に最も対応する候補を選択する方法について述べる．

2 非対訳コーパスからの対訳獲得

対訳関係のないコーパスから対訳表現を獲得する場合には，対訳コーパスを使う場合と違い，単語の出現位置や頻度情報を直接利用することができない．Fungは，対訳となる語はそれぞれのコーパスで共起する語が類似すると仮定し，中国語と英語のコンパラブルコーパスから対訳表現を獲得している [1]．この方法では各未知語について共起語のベクトルをつくり，類似性の高い組み合わせを対訳語として判定している．各語について対訳となる候補の数が多くなると全ての組み合わせについて類似性を評価するのが困難になり，ある程度比較する候補を絞り込んでおく必要がある．

本稿では，専門用語の日本語と英語の対訳を獲得することを目的として日英の複合名詞の意味的対応を判定する方法について述べる．複合名詞の対訳では，構成語の対応関係を利用して比較する訳語候補を絞り込んでおくことができる [3]．以下の手順により，与えられた日本語の複合名詞 c_J の英訳語 c_E を検索する．

1. 品詞パターンにより英語コーパスから複合名詞候補集合 C_E を収集する
2. 対訳辞書，シソーラスを用いて C_E から c_J と構成語の対応がとれる語を訳語候補集合 T_E として抽出する
3. T_E から c_J の訳語としてもっともらしい語 c_E を選択する

はじめに1で，品詞パターンを用いて英語コーパスから獲得対象とする表現 C_E を収集する．例えば，典型的な複合名詞の品詞構成である「名詞+名詞」や「形容詞+名詞」の単語列を集める．次に2で，もとの日本語と辞書やシソーラスを用いて構成語の対応のとれる表現集合 T_E を抽出する．例えば c_J = “販売効率” の場合は，T_E の要素として “sales productivity”, “sales rate”, “market efficiency” などが抽出される．3で，これらの候補から共起語の類似性を調べて，もっともらしい訳語候補を選択する．出現頻度の最も高い語を選択することもできるが，もとの日本語や訳語候補が使われている文脈を考慮していないため精度に限界がある．本稿では，2で集められた候補から3で共起語の情報を使って

もとの語と意味的に最も対応する c_E を選択する方法について述べる．

3 共起語と共起意味属性

異なる複合名詞の類似性を調べるために，該当する語と同一文に共起する語を使う．共起語の対象語との関係は，統語的な依存関係の有無によって，次の「生産能力」とその共起語の例のように分けることができる．

1. 依存関係あり
   - 格関係
     例 [生産能力] 拡大
     [生産能力] を 強化する
   - 連体修飾関係
     例 余剰な [生産能力]
2. 依存関係なし
   - 比較/対比
     例 現在は 需要 が [生産能力] を上回っている．
   - 連想/関連
     例 新 工場 が完成すれば全 [生産能力] は過去最大となる．
     最新鋭 設備 を導入して [生産能力] を三倍に引き上げる．

1 と 2 はどちらも対象語「生産能力」の特徴を表す語と考えられるが，その性格は異なるため両者を区別して扱えると都合が良い．1 の単語は主辞（「能力」）との関連性が強く，異なる主辞をもつ複合名詞を区別するには良いが逆に同一主辞の複合名詞の判別能力は低くなると考えられる．

[生産能力] ｝拡大，強化する，不足する，
[輸送能力] ｝過剰な，適正な

一方 2 には，修飾語（「生産」），あるいは複合名詞全体に関連の深い語が含まれると考えられる．

[生産能力]： 需要，工場，メモリ
[輸送能力]： 貨物，トラック，バス

構文解析を行えば共起語がどちらに分類されるかを決定することができるが，現状では時間的コストに見合う解析精度を得ることは難しい．そこで，本稿では次のように共起語をその品詞で分類し，各分類ごとに特徴を調べる．

1. 動詞
2. 用言性名詞
3. 2 以外の一般名詞

1 と 2 を依存関係のある共起語，3 を依存関係のない共起語に近似して用いる．大雑把な方法ではあるが，ある程度の頻度で共起する語に対しては適当な分類になると期待される．

複合名詞 c とその共起語 r の関連性の強さの指標として相互情報量 MI[4] を用い，r の特徴量 $\lambda_c(r)$ を次のように定義する．

$$\lambda_c(r) = \begin{cases} MI(c,r) & : f(c,r) \neq 0 \\ 0 & : f(c,r) = 0 \end{cases} \quad (1)$$

$$MI(c,r) = \log \frac{f(c,r)F}{f(c)f(r)} \quad (2)$$

ここで，$f(c)$，$f(r)$ はそれぞれ複合名詞 c，共起語 r の出現頻度，$f(c,r)$ は，c，r の共起頻度，F はコーパス中の全単語の総出現頻度である．

また，データのスパースネスを補完するため，共起語をシソーラスを使って意味属性に抽象化し特徴量を計算することも行う．本稿ではシソーラスとして日本語語彙大系 [5] を使った．語彙大系は，日本語の名詞に約 2,700 の意味属性を付与しており，共起語をこの属性に置き換えて日英間で共起する意味属性を比較した．一つの表記で表された語が複数の意味属性を持つことがあるので，次のようにして一つの属性 a を選び，単語の場合に準じて $\lambda_c(a)$ を計算する．

1. 共起語がもつ全ての意味属性の頻度を集計する
2. 各共起語について，最も高い頻度をもつ意味属性をその語の意味属性として採用し，全ての共起語を一つの意味属性に対応させる
3. 決定された意味属性について再度集計する

例えば，「交通機関」の共起語「バス」が “乗り物”，“部屋” という二つの意味属性を持っていた場合，「交通機関」の全ての共起語の意味属性を集計した結果，“乗り物” の頻度が “部屋” の頻度を上回っていれば，「バス」の意味属性として “乗り物” を採用する．その後再度全ての共起語を選択された意味属性に変換し特徴量の計算を行う．英語に関しては，対訳辞書で日本語に変換しそれぞれの日本語に付与されている意味属性をもとの英語の属性と考えて同様の処理を行う．

営業利益		operating profit		business interest	
[用言性名詞]					
節減	9.15			economy	5.52
低減	9.13	fall	10.71		
配当	8.88	share	6.81		
響き	8.80			sound	9.01
予想	8.48	expectation	12.27		
[一般名詞]					
歩留まり	9.99	yield	6.60		
工賃	9.79	charge	8.48	charge	7.29
利益	9.43	profit	7.81		
経費	8.61	expense	7.55	expense	10.49
採算	8.36	profit	7.81		
[動詞]					
減る	9.44			diminish	11.54
算入する	9.39	include	7.60	include	9.06
増える	8.69	increase	9.29		
倍増する	8.16	double	8.22		
[意味属性]					
“費用”	19.89		15.96		14.43
“利益”	19.86		16.92		
“予想”	18.92		18.07		
“収入”	18.84		15.52		
“増加”	18.70		15.12		
“期間”	18.60		15.30		13.20

図 1: 複合名詞の共起語，共起意味属性の比較

4 日英複合名詞の文脈による比較

3節で述べたように共起語あるいは共起意味属性を使って対象とする複合名詞の特徴を表し，その類似性により日英の複合名詞の対応度を調べる．言語が異なる共起語を対応付けるために，対訳辞書の単語を利用する．ただし,「居る」,「所」のように語義の多い語を対応付けに用いると比較に悪影響を与えると考えられるので対訳辞書において対応する訳語数が5をこえる語は用いないこととする．図1は,「営業利益」と “operating profit”, “business interest” のそれぞれの共起語のうち対訳辞書で対応のとれたものを品詞別に分けて日本語共起語の相互情報量の高いものから順に示したものである．用言性名詞，動詞は「生産能力」と格関係を持ちやすい語が上位に来ていることがわかる．また，一般名詞では関連する語として共起する語（歩留まり，工賃，利益）が上位に現れている．

英語訳候補の共起語との対応をとると，いずれの分類でも “operating profit” の方が，“business interest” よりも相互情報量の高い語に対応のとれるものが多いことがわかる．日英の複合名詞の類似度を定量的に評価するために，日英複合名詞 c_J ， c_E の類似性の評価値を次のように定義する．それぞれの共起語（品詞 p）r_{Ji}, r_{Ei} の特徴量 $\lambda_{c_J}(r_{Ji}), \lambda_{c_E}(r_{Ei})$

	高頻度語	低頻度語
共起語評価値 $S_w(1)$	71	61
共起語評価値 $S_w(2)$	66	57
共起意味属性評価値 S_a	61	66
最頻度	63	53

表 1: 候補選択の正解数

を要素とする複合名詞の特徴ベクトル

$$\boldsymbol{v}_{Jp} = (\lambda_{c_J}(r_{J1}), ..., \lambda_{c_J}(r_{Jn})) \quad (3)$$

$$\boldsymbol{v}_{Ep} = (\lambda_{c_E}(r_{E1}), ..., \lambda_{c_E}(r_{En})) \quad (4)$$

を考える．ここで，r_{Ji} と r_{Ei} は対訳辞書により対応付けられた組である．その内積を品詞ごとに重み $w(p)$ を乗じ足し合わせた $S_w(c_J, c_E)$ を共起語評価値として使う．

$$S_w(c_J, c_E) = \sum_p w(p)\boldsymbol{v}_{Jp}\boldsymbol{v}_{Ep} \quad (5)$$

意味属性に関する評価値 S_a も意味属性の特徴量 $\lambda_{c_J}(a_i)$, $\lambda_{c_E}(a_i)$ を要素とするベクトル使って同様に定義する．

5 実験と考察

コーパスとして日本経済新聞 CD-ROM95 版，Wall Street Journal1996 年版を用いて日本語の複合名詞の対訳となる英語を検索する実験を行った．コーパス中に出現する日本語複合名詞を高頻度語

	英訳語候補	$S_w(1)$	S_a
技術移転	technology transfer	2101	4221
	technology share	1345	1478
	policy move	807	2311
生産能力	production capacity	5748	5830
	output capacity	1357	5245
	production capability	1283	2235
情報	information system	15661	16533
システム	data system	15437	22220
営業利益	operating profit	4934	8706
	business interest	2236	4845
	trading profit	1666	4712
証券会社	securities company	4711	5829
	paper company	4087	8159
電力会社	energy company	6125	12909
	power company	5637	9327

図 2: 訳語候補の例

(頻度 100 以上) と低頻度語 (頻度 100 未満) に分け，2 節で述べたように構成語の対応を使って訳語候補 T_E を集めた．これらの日本語から T_E の中に正解訳が含まれるものを 100 語ずつ選び，もとの日本語との共起語評価値 S_w あるいは共起意味属性評価値 S_a が最も高い候補を英語訳として選択した，S_w の品詞種類による重み $w(p)$ は，(1) w(動詞) $= 0$, w(用言性名詞) $= 0.5$, w(一般名詞) $= 1$, (2) w(動詞) $= 1$, w(用言性名詞) $= 0.5$, w(一般名詞) $= 0$ の 2 種類で行った．これらを T_E の中で最も出現頻度の高い候補を選んだ場合の結果と比較した．

表 1 にその結果を示す．共起語評価値を使用し重みが (1) のとき，高頻度語については単純に高い頻度の候補を選択した場合に比べて正解率が数ポイント向上している．また，共起語の品詞について比較すると，一般名詞を使った (1) の方が動詞を使った (2) より高い値が出ており，本実験のように主辞が同じものを含むような候補を選択する場合には動詞よりも，関連/連想語を含む一般名詞の方が判別に寄与している結果となった．

共起意味属性評価値を用いた場合，高頻度語については低い値であるが，低頻度語については最も良い結果となっている．これは，低頻度語では日英で対応付けられる共起語が少なくなり共起語評価値での精度が落ちるが，意味属性に抽象化することによってそれが補われた効果であると考えられる．

図 2 は訳語候補の例である．コーパスから英語訳候補を抽出する際には構成語の対応を広くとっているので複数の候補が現れているが，概ね評価値の上位の候補が適切な英語訳となっている．主辞が異なるもの (“technology transfer”, “technology share”)，主辞が同一のもの (“securities company”, “paper company”) どちらも共起語評価値により適切に分別できている．しかし，“energy company” と”power company” のように包含関係にあるものは選択に失敗している．他の例として「輸出価格」の訳語に “export price” より “import price”[1]を選択したものがある．これらのように上位下位関係になる語や対となる語は同じような文脈で使用されることが多く，本手法のみでこれらを区別することは困難である．実験では文脈による選択効果を見るために全ての訳語候補を同等に扱っているが，構成語の対応が対訳辞書によるものかシソーラスによるものかという情報を使えばこの問題は回避できる．

6 おわりに

共起語，共起意味属性を用いて複合名詞の意味的対応を判定する方法を提案し，非対訳コーパスからの複合名詞対訳の獲得における訳語候補の選択に有効であることを確認した．本稿の実験のように主辞が同一になるような類似の複合名詞から適切な訳語を選択する場合には一般名詞の重みを動詞，用言性名詞より高くする方が精度が良い結果となったがタスクによる最適な重み付けについては検討が必要である．共起語/意味属性による評価値は訳語の選択以外にも似た意味の語のグループへのクラスタリングに利用することも考えられる．

参考文献

[1] P. Fung, “A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora,” Lecture Notes Computer Science, vol. 1529, pp. 1-17, 1998

[2] K. Tanaka and H. Iwasaki, “Extraction of lexical translation from non-aligned corpora,” Proc. of 16th COLING, vol. 2, pp. 580-585, Aug., 1996

[3] T. Tanaka and Y. Matsuo, “Extraction of translation equivalents from Non-Parallel Corpora,” Proc. of 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI99), pp. 109-119, Aug., 1999

[4] K. W. Church and P. Hanks, “Word association norms, mutual information and lexicography,” Proc. of 26th ACL, pp. 76-83, 1989

[5] 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 (編)，日本語語彙大系，岩波書店，1997

[1]「輸出」と “import” がシソーラスにより対応付けられたため訳語候補として挙っている．