形容詞－名詞組の日英対訳獲得

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

1 はじめに

自然言語処理では表層的な表現と表される意味内容との対応をとることが大きな問題である．個々の意味内容を定義すること自体簡単な問題ではないが，複数の語を組み合わせることで表層表現の表す意味内容の範囲を限定することができる．語の組み合わせとしては，同一言語内で関連のある語の組，異なる言語で同一の意味を表す語の組が考えられる．

同一言語内で関連性の強い語には形容詞－名詞の組み合わせがある．ある概念を表す名詞には様々な形容詞，形容動詞[1]などの連体修飾語を付加することができるが制約なくあらゆる語が使えるわけではない．「暖かい天気」「良い天気」という表現はよく使うが「元気な天気」「高い天気」という表現はあまり使われない．「天気」という語が表している概念に「元気だ」「高い」という語に形容される属性がないためであると考えることができるが，このことは共起する形容詞が名詞の表している概念の特徴 (属性) を間接的に説明していると見ることができる．

また，一つの言語の中で多義を持つ表現であっても別の言語の表現と対応させることにより表される意味内容との対応の曖昧性を減少させることができる．例えば，英語の「plant」には多義があるが，日本語と対にして「plant-植物」「plant-工場」などとすることで表される意味内容が限定される．

この両者の組み合わせたもの，つまり形容詞－名詞の組の対訳表現はある程度限定された意味内容を表すと考えられる．本稿では，連体修飾関係にある形容詞－名詞組を単位として考え，コーパスから日本語と英語の対訳を収集する方法について検討する．また，本手法で収集された対訳の特徴について述べる．

2 形容詞と名詞

西尾らは形容詞をその表す意味により感情形容詞と属性形容詞に分類している [1]．感情形容詞は主体の感情を表す形容詞（楽しい，苦しい），属性形容詞は修飾される対象の属性を表す形容詞（優しい，高い，難しい）である．また，内海らは形容詞の多義性について分析し，第一語義とその他の語義の関係について述べている [2]．その中で感情形容詞は共通の特徴として主体の感情の側面を失って被修飾対象の客観的特徴を表す属性形容詞として働くことを指摘している[2]．

N1(具体物) が 濃い	–	N1 be thick
N1(色彩) が 濃い	–	N1 be dark
N1(茶 炊事 酒) が 濃い	–	N1 be strong
N1(関係) が 濃い	–	N1 be close
N1(霧) が 濃い	–	N1 be dense

図 1: 形容詞対訳の例（「日本語語彙大系」[3] より）

例： 曲調が楽しい，経営が苦しい

本稿では属性形容詞が名詞を連体修飾する表現を対象とするが，この場合第一語義が感情形容詞であっても属性形容詞の用法で使用される場合が多いと考えられるので特にこれらを区別することはしない．

例： 楽しい話，苦しい時代

ある名詞によく共起する形容詞はその名詞の特徴的な属性を修飾していると考えられるが，日本語と英語のように文化的背景の大きく異なった言語間では同一の概念の属性を修飾するのに用いられる語が異なることがよく知られている．例えば，「濃いコーヒー」という表現は「強いコーヒー」よりも普通に使われているが，逆に “thick coffee” や “deep coffee” よりも “strong coffee” という表現の方がよく使われる．どの形容詞が選択されるかは被修飾名詞に密接に関係しており，日本語と英語では一対一に対応しないことが多い (図 1)．

しかし，同様の概念を表す名詞同士を比較するとやはりそれらによく共起する形容詞も類似した概念を修飾するものが現われると考えられる．日本語と英語間で対訳となる名詞それぞれに共起する形容詞

[1] 本稿では形容動詞も形容詞と同様に扱う

[2] 西尾らはこれらを別の語義とは区別していない．

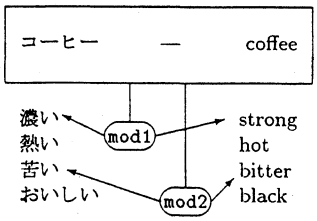

図 2: 形容詞共起語の対応

を対応付けることにより同様の意味内容を表す形容詞ー名詞の組を獲得できると期待される (図 2).

3 コーパスからの形容詞ー名詞の対訳獲得

コーパスから対訳表現を獲得するには，対訳コーパスを用いる方法が一般的である．しかし，対訳が用意されているコーパスに比べてそうでないものの方が圧倒的に多いので，単言語のコーパスを利用して対訳表現を獲得しようとする研究も行われている [4, 5, 6, 7].

Dagan らは *subject-verb*, *verb-object* など目的言語コーパス中で統語的依存関係にある単語の組合わせの情報を利用して原言語の語義の曖昧性解消を行っている [4]．また田中らは名詞 2 語連続などの品詞列を利用して対訳関係のないコーパスから複合名詞の対訳を獲得する方法を提案している [7]．統語的関係がある語の組み合わせは，異なる言語においても同様の関係を持っている場合が多いので対訳を探索する手がかりとなる．本稿では，日本語，英語それぞれについて連体修飾関係をもつ形容詞ー名詞組をコーパスから収集し，それらを単語単位の対訳辞書を用いて両言語の対応付けを行う．しかし，2 節でも述べたように日本語と英語では文化的背景が大きく異なり，単語の表す語義の対応関係は多様であるため通常の対訳辞書のみでは適切な候補を対応付けられない可能性がある．そこで日本語と対応付ける英語の候補をシソーラスを用いてその類義語まで広げる．ただし，形容詞，名詞双方の類義語を考慮すると不適切な候補が多くなるので，名詞は対訳辞書にある組み合わせのみを対応付け，形容詞はそれに加えてシソーラスによる候補の拡張を行う．

以上をまとめると次の手がかりを利用してコーパスに出現する二言語間の表現を対応付ける．

{healthy, salubrious} ←&→
{wholesome} ←&→
{hearty, satisfying, solid, substantial}

図 3: 形容詞の synset 間のリンク

1. 異なる言語間の対応
   - 対訳辞書（名詞，形容詞の対応）
2. 同言語内の対応
   - シソーラス（形容詞の類義語）
   - 統語的依存関係 (形容詞ー名詞の連体修飾関係)

4 形容詞の類義語の収集

英語の形容詞の類義語を収集するのに英語の語彙データベース WordNet[8] を利用する．WordNet では同義語の集合 synset によって一つの概念を表しており synset 間に上位/下位関係，全体/部分関係などのリンクが張られている．WordNet は，名詞，動詞，形容詞，副詞の 4 つの辞書から構成されており，そのうち形容詞辞書中の synset 間には，類似 (similar) を表すリンクが張られている．図 3 は形容詞の synset の例で，{...} は synset を，←&→ は類似リンクを表している．本稿では，形容詞間の類似度をこのリンクを辿る距離によって定義する．類似リンクの距離を 1，同一 synset 内にある単語間の距離を 0.5 と定義する．例では，*healthy* と *salubrious* の距離は 0.5，*healthy* と *solid* の距離は 2 である．通常一つの字面は複数の synset に含まれるので 2 単語間の経路は複数あるがそのうちで最短の距離を採用する．本稿では，距離が 2 以下の表現を類義語として扱った．

5 実験

コーパスから収集した日本語と英語の形容詞ー名詞組を前節で述べた方法を用いて対応付けをする実験を行った．コーパスは日本経済新聞 CD-ROM94 年版 (NIK，約 4,300 万語) と Wall Street Journal 96 年版 (WSJ，約 3,100 万語) を使用し，辞書は 日英機械翻訳システム ALT-J/E[9] の対訳辞書（一般名詞約 83,000 語，形容詞約 24,000 語）を使った．NIK，WSJ から隣接して共起する形容詞と名詞の組を収集した．ただし，単に隣接する形容詞と名詞を取ると依存関係のない 2 語を取る可能性があるの

日本語： [ˆ の]**Adj N** [ˆN Suffix の]
英語： **Adj N** [ˆN]
[ˆ] は [ ] 内の以外の単語を表す

図 4: 形容詞ー名詞のテンプレート

日本語	英語	dist	freq	jud
独創的な技術	creative technology	0.0	2	o
	original technology	0.0	1	o
優秀な技術	superior technology	0.0	11	o
	great technology	0.5	2	o
根強い需要	strong demand	0.5	304	o
	solid demand	0.5	2	o
巨大な需要	huge demand	0.0	20	o
	enormous demand	0.0	2	o
公正な社会	good society	0.5	4	<
	open society	1.0	16	<
豊かな社会	rich society	0.0	1	o
	prosperous society	0.5	3	o

表 1: 収集表現の例

で図 4 のテンプレートにあてはまるような，直前や直後に助詞や他の名詞や接辞などが現われていない 2 語を収集した．

3 節で述べたように，対訳辞書と WordNet を用いて日本語と英語の形容詞ー名詞組の対応付けを行った．出来た組の中から NIK の中で高頻度で出現する名詞 26 種類が含まれるものを選び，これらの対訳としての妥当性評価した．評価対象となった対訳の日本語の種類は 60 で，日本語一つにつき英語訳候補を最大 2 種類取り出したので対訳数は 117 組である．同一の日本語の中で英語は，4 節で定義した形容詞間の距離が短いもの，出現頻度の大きいものの順で順位付けをしている．獲得された対訳の例を表 1 に示す．

評価は，対象の日本語，英語表現を含む文をそれぞれ NIK,WSJ から最大 3 文ずつ抜き出し[3]以下の 4 種類の判定をした．

適切 (o) 文脈を考慮して意味の一致する文が 1 組以上ある

	本手法 (一位)	本手法 (二位まで)	MT
適切	34 (56.7%)	36 (60.0%)	45 (75.0%)
J>E	5 (8.3%)	4 (6.7%)	3 (5.0%)
J<E	4 (6.7%)	4 (6.7%)	0 (0.0%)
不適切	17 (29.3%)	16 (26.7%)	12 (20.0%)

表 2: 対訳収集結果

[3]実際には文脈を明確にするため各文の直前の 1 文を含む．

収集表現の方が適切	10
同等 (良)	28
同等 (悪)	4
MT の方が適切	11
合計	53

表 3: 収集表現と MT の差分

J>E (>) 例文中の文脈では意味が異なるが，一致する文脈が考えられる（日本語の意味が広い）

J<E (<) 例文中の文脈では意味が異なるが，一致する文脈が考えられる（英語の意味が広い）

不適切 (×) 対訳として不適切である．

また，比較のため対象となった日本語を機械翻訳システム ALT-J/E (以下 MT) で英語に変換した．ただし，文脈は翻訳結果がコーパス中にあるもののみ考慮した．それぞれの妥当性を日本語を単位として調べた結果を表 2 に示す．

提案手法も MT と同じ対訳辞書を使用しているにもかかわらず適切な対訳と判断された数は MT に及んでいない．これは適切とされた表現が多くがコーパス中に出現していなかったため，WordNet で拡張した類義語を引いた結果不適切な表現を選択してしまったのが原因である．ただし，文脈を考慮すれば対訳として使える組み合わせ（J>E，E<J）を含めると，提案手法（2 位まで）で 73.4%，MT が 80%で差は小さくなる．

提案手法はコーパス中にある表現しか収集できないので，得られた表現の多くが MT の結果と異なるものなった．その差分の内訳を表 3 に示す．MT と比べて収集した表現の方が良いものと同程度に良いものを合わせて差分が出た表現の過半数になっている．このことは，実際の使用される形容詞表現が対訳辞書中の候補より広く多様であることを示唆している．

6 考察

6.1 誤りの分析

前節で述べたように誤った対訳を収集した最も大きな原因は正解とする表現がコーパスに出現しなかったことであるが他に以下の点が挙げられる．

1. 不適切な類義語を選択している
2. 適切な候補がない

1 は WordNet によって不適切な類義語と対応させてしまった場合である．

例 大胆な政策 – original policy

この例では，「大胆な」–*daring*→{*daring, avant-garde*} ←&→{*original*} という経路を辿って類義語を得ているが，元の「大胆な」の意味とずれが生じている．

2 は対訳辞書に収録されてなかった対応が現れた場合である．

例 冷ややかな声 – harsh voice

適切な英語訳としては *negative comment* や *critical comment* などが考えられるが，「声」–*comment* という対訳が辞書になかったため候補に挙がっていない．この種の対訳は翻訳の観点からも辞書に追加する必要があると思われる．

6.2 収集表現の分析

収集された結果を観察し，訳語選択，換言処理との関連について述べる．

日本語ー英語の組み合わせにより意味が限定

- 優秀な技術 – superior technology
  cf. 優秀な技術 – superior technique

日本語の「技術」は "techonology"(科学技術) の意味と "technique"(技巧) の意味を含んでいるがどちらも形容動詞では区別できない．訳し分けるには「技術」の内容の解釈が必要になる．

- 厳しい環境 – harsh environment
- 厳しい環境 – dangerous environment

「厳しい環境」には「過酷」(*harsh*) な場合と「危険」(*dangerous*) な場合が含まれると考えられるが，区別するためには「厳しい」の内容を捉えなければならず難しい．

形容詞ー名詞の組み合わせにより意味が限定

- 強硬な姿勢 – strong stance
  cf. 正しい姿勢 – good posture

修飾する形容詞「強硬な」に対して，「姿勢」の内容を態度の意味 (*stance*) ととり体の姿勢の意味 (*posture*) と区別される．

類似した表現と対応付けられた例

- 根強い需要 – solid demand

「根強い」と *solid* で直接の類似性は高くないが表現全体で類似性をもっている．*solid* を日本語に逆引きすることで「堅実な需要」という換言表現を得ることができる．

限定された意味の表現と対応付けられた例

- 弱い大統領 – spineless president

特定の文脈でより限定された意味の形容詞と結び付けられることがある．*spineless* というを狭い意味の表現を「弱い」に対応付けることでより一般的な意味での換言が可能になる．

7 おわりに

対訳辞書とシソーラスを使うことによりコーパスから連体修飾関係にある形容詞ー名詞組の対訳を獲得できることを示した．形容詞は修飾する名詞との関係によって多様な語が使われるがシソーラスを用いることで類似した意味を持つ形容詞ー名詞表現を柔軟に対応付けることができる．また，形容詞ー名詞のまとまった表現を異なった言語間で対応させることによって意味内容を限定したり同一言語内の換言処理に利用することも可能であることを示した．

参考文献

[1] 西尾寅弥：形容詞の意味用法の記述的研究，秀英出版，1972

[2] 内海彰，堀浩一，大須賀節雄：自然言語処理のための形容詞の意味表現，人工知能学会誌，8-2，pp.192-200，1993

[3] 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

[4] I. Dagan & A. Itai: Word Sense Disambiguation Using a Second Language Monolingual Corpus, *Computational Linguistics*, 20(4), pp. 563-596, 1994

[5] 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

[6] 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, 1999

[7] 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表現の獲得，電子情報通信学会論文誌，Vol.J84-D-II，No.12，pp.2605-2614

[8] C. Fellbaum (ed.): WordNet: An Electronic Lexical Database, MIT Press, 1997

[9] Satoru Ikehara: Multi-level Machine Translation Method, *Journal of Future Computing Systems*, Vol.2, No.3, 1989