辞書情報と共起情報による日英翻訳用複合名詞解析

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

複合名詞の処理は自然言語処理全般に関わる大きな問題である。語基となる名詞や接辞の組み合わせによって無数に新しい語がつくり出されるため、辞書に複合名詞全てを網羅することは事実上不可能である。そこで、その構成している名詞間のつながりに着目し、内部構造を解析する研究が数多く行われている。

しかし、実際に使用される複合名詞は使用される頻度やその経過によって個々の語基間の結束性が変化することも多く、人間が見ても正しい解析結果を一意に決定することは簡単なことではない。

そこで、我々は日本語の複合名詞の解析方法について、解析結果を機械翻訳システムに適用し英語に変換するということを観点として考察した。

2 従来手法と問題点

複合名詞の係り受け構造を決定するのには、コーパス中の統計情報を利用して構成語間の関連度や結束性を調べる方法が提案されている [1][2][3]。しかし、出現頻度、相互情報量などで単語間の結び付きを計測しただけではうまくいかない場合も少なくない。例えば、図1は複合名詞中の隣接する2語について新聞1年分中の bigram で相互情報量を測定した結果である。(1) の「希望小売価格」という語では「小売価格」を部分複合名詞として扱い (希望,(小売, 価格)) と解析する方が自然と思われるが、「価格」という語の出現頻度が「希望」「小売」に比較してかなり高いため相互情報量が相対的に低くなり、(小売, 価格) の結び付きを (希望, 小売) より弱いという判断してしまう。また、(2) の例では (土地,(区画, 整理)) という本来の意味「[土地の区画] を [整理する] こと」とは異なった結果を導いてしまう。これは、「土地区画整理」という語のなかの「区画」はわざわざ明示しなくても「土地」の「区画」であることが分かるため「区画整理」という語で同様の内容を表すことができるからと考えられる。その結果コーパス中で「区画整理」の頻度が高くなり、他文脈ではあまり現れなかった「土地区画」が部分複合名詞として認められないからと考えられる。このようにコーパス中の統計量から得られた構成単語間の結び付きの強さと、意味的な構造は必ずしも一致しない。

(1)	希望	小売	価格
(MI)	11.0	8.3	
(2)	土地	区画	整理
(MI)	10.1	13.0	

図 1: 相互情報量による構造解析

日英機械翻訳を前提とした場合に望ましい解析という観点で見ると (1),(2) には違いがある。(1) については、「小売価格」という複合名詞に対して “retail price” という英語訳が存在しており、やはり「希望」＋「小売価格」＝ “desirable” ＋ “retail price” という解釈が妥当であると考えられる。

これに対し、(2) では英語訳 “land reallocation” との対応関係を見ると「土地」＋「区画整理」＝“land”＋“reallocation” の解釈の方が英語の変換には都合が良い。つまり「区画整理」の語は一語としての性格が強く、「[土地] を [区画整理する] こと」という解釈が成り立つと考えられる。

日英翻訳の観点からみると、このように英語との対応付けがしやすい構造を考えると都合が良い。そこで、「区画整理」のように単独で英語に変換できるような複合名詞 (部分複合名詞とよぶ) を他の語基と同様に扱い、これらの間で構造を生成することを考える。

本稿では、日本語の複合名詞を始めにコーパス中の共起情報を利用して意味的なまとまりを持つ部分複合名詞に分割し、その後語基や部分複合名詞間についての係り受け関係を調べ、構造解析を行う方法について述べる。

3 複合名詞の種類と部分複合名詞

本稿では扱う複合名詞は、「辞書記載された名詞の列で全体として文法的に名詞として振る舞うもの」[3] という定義に接辞を加え、接辞と名詞の列からなる表現を対象とする。この定義にあてはまる複合名詞はいくつかの種類に分けられる。

1. 固有名詞
   (例) 日本電信電話株式会社、国防総省
2. 専門用語: ある分野で固定的に使用される語
   (例) 経常利益、景気動向指数
3. 事象: 末尾が用言性名詞で文に展開できる語
   (例) 規制強化、消費税廃止、米穀小売価格調査
4. その他一般
   (例) 携帯電話、基本方針

複合名詞をこれらに分類することは必ずしも容易ではないが、内部の語の結び付きの強さ、性格は異なっており、同じ係り受け構造として記述するのはあまり得策ではないと考えられる。例えば、1. の内部構造を細かく厳密に解析する必要はないが、3. の語は用言性名詞に対する係り方を解析する必要がある。

また、3. の場合、複合名詞の中に 2. や 4. で見られるような単独で用いられる複合名詞が含まれている場合があり、これらを無視して一様に展開するよりも、含まれている複合名詞をひとかたまりとする方が扱いやすいと考えられる。例えば、「米穀/小売/価格/調査」を「(((米穀ヲ小売スル) 価格ヲ) 調査スル)」と展開するよりも、「(((米穀ノ) 小売価格ヲ) 調査スル)」とした方が "survey on retail price of rice" と変換しやすい。このような複合名詞内に含まれるより小さな複合名詞 (部分複合名詞) や語基が結合してより大きな複合名詞を形成していくとする考え方は宮崎らが述べており、数詞や固有名詞、これらと接辞、用言性名詞との結合の仕方などに注目してルールを作成し、部分複合名詞を認定している [4]。しかし、適用する複合名詞の使われている分野の特徴に応じてルールを人手で修正したり拡充することは容易ではない。

そこで、本稿ではコーパスと辞書の情報を利用して、単語に分かち書きされた複合名詞の係り受け構造を決定する方法を検討する。始めにコーパス中に含まれる複合名詞表現の情報からその使用頻度や他の語と結合した使われ方を調べ、一語として扱う部分複合名詞を決定する。その後、文法規則と格フレーム辞書を使って語基と部分複合名詞の係り受け構造を生成し、コーパス中の統計情報を用いて尤度を計算して、最終的な構造を決定する。

4 複合名詞の構造

本稿では、複合名詞の構造を文献 [3] で述べられているような二分木の形で表現する。これは、図 2 のように括弧を用いて 2 項関係として表現できる。日本語の場合名詞 2 語からなる複合名詞では通常右側にある語が主辞となるので、括弧でくくられた表現の主辞は、括弧内で最も右側にある語とする。また、後述するように独立性の高い部分構造を部分複合名詞として認定し、"@(,)" で表して他と区別する。

((景気, 動向), 指数)
((簡易, 型),@(携帯, 電話))

図 2: 複合名詞の構造

5 複合名詞の部分分割

解析対象となる複合名詞表現を、内部に含まれる短い複合名詞 (部分複合名詞) に分割することを考える。ここで、分割対象とする部分複合名詞として以下のような表現を考える。

- 単独で使用される頻度が高い表現
- 結合してより長い複合名詞の一部となる場合にはその種類が多様である表現

これらの条件を満たす表現は図 3 の「携帯電話」の例のようなもので、独立した形で現れ多様な複合名詞を生成するので部分複合名詞として認められる。一方「電話システム」は単独で用いられず、他の複合名詞の一部として現れるのみであるので部分複合名詞とはしない。この条件を判定するのに専門用語の抽出などに利用される *C-value*((1) 式) を用いた [5]。

$$C\text{-}value(a) = n(a) - \frac{t(a)}{c(a)} \qquad (1)$$

ここで、a が語基列、$n(a)$ は語基列 a が現れる総頻度、$t(a)$ は、a を含むより長い表現 xay が現れる頻度、$c(a)$ は xay の異なり数である。この *C-value*

が閾値をこえる表現を部分複合名詞とする。出現頻度が高い表現であっても、より長い複合名詞の一部としてしか出現しない表現は、*C-value* の値が小さくなり独立性の高い部分複合名詞とは認定されない。

各複合名詞について、隣接する語基あるいは部分複合名詞の組ごとにそれらが結合した表現が部分複合名詞と認められるかを判定し、これ以上まとめられなくなるまで結合を繰り返す。その結果、最終的に認定された部分複合名詞も二分木の形で表される (図 4)。

			頻度
	携帯/電話		79
	携帯/電話	/市場	12
簡易型/	携帯/電話		7
	携帯/電話	/会社	3
デジタル/	携帯/電話		2
次世代/	携帯/電話		1
	...		...
	(総頻度)		134
	電話/システム		0
移動/	電話/システム		2
携帯/	電話/システム		1
自動車/	電話/システム		1
公衆/	電話/システム		1
	(総頻度)		6

図 3: 部分複合名詞の判定

制限, 付き,@(一般,@(競争, 入札))

図 4: 部分複合名詞

6 係り受け構造の生成

分かち書きされた語基および部分複合名詞から簡単な CFG ルールによって構造を生成する。係り受けの種類は、格修飾による関係とそれ以外の修飾関係の 2 種類のみを考える。複合名詞に用言性の名詞を含む場合には、格フレームを利用して係り受け構造を生成する。図 5 のように妥当な格フレームの適用の仕方が複数ある場合には、全ての構造を生成し、後に行うコーパス中共起情報による尤度計算によって曖昧性を解消する。

7 係り受け構造の尤度計算

6 で生成した各係り受け構造に対して、格フレームとコーパス中の共起情報に基づいて尤度計算を

(1) (政府 が 開発) する (の) を 援助 する

(2) 政府 が (開発 を 援助) する

図 5: 格フレームによる構造解析

政治/N 改革/N 推進/N 派/S_x
→ (政治, 改革) (政治, 推進) (政治, 派)
(改革, 推進) (改革, 派) (推進, 派)

図 6: 複合名詞中共起

行う。

7.1 コーパス中の共起情報

係り受け尤度の計算や部分複合名詞を判定するために、コーパス中での各語基の共起情報を用いる。解析対象が複合名詞であるので、コーパス中に現れる複合名詞を単位として共起情報をとる。統計量を簡単に獲得するため、以下の品詞列の並びを複合名詞としてみなす。

$$[P_x, N][P_x, N, S_x]*[N, S_x]$$

ここで記号 P_x、N、S_x はそれぞれ接頭辞、名詞、接尾辞を、"[ ]" は中に記述された品詞からなる品詞クラス、"*" は 0 回以上の繰り返しを表している。これを複合名詞列と呼ぶ。

この複合名詞列のなかの全ての 2 語基の組合わせを共起した語基として取出し、共起頻度を計数する (図 6)。また共起情報は、前の語基から後ろの語基に係る確からしさに用いるため語基の順序は保持する。つまり (政治, 改革) と (改革, 政治) は別の共起として扱う。

7.2 2 項間係り受け尤度の計算

生成した係り受け構造について、その妥当性を計算する。コーパス中の共起情報を用いて各語基あるいは部分複合名詞の 2 項関係の結合の強さを判定する。部分複合名詞については、その語の主辞となる語基で代表させる。

本稿では語基 x,y の結合の強さとして語順を考慮した相互情報量 $I(x,y)$((2) 式) を用いた[2]。

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x,*)P(*,y)} \qquad (2)$$

ここで、$P(x,*)$,$P(*,y)$ はそれぞれ語 x が複合名詞中で末尾以外に出現する確率, y が先頭以外に出

現する確率、$P(x,y)$ は複合名詞内で語 x の後に語 y が共起する確率である。つまり、$P(x,y)$ と $P(y,x)$ は区別する。出現確率、共起確率は、コーパスから 7.1 で述べた複合名詞列を収集しその中での出現頻度から計算する。

生成された各複合名詞構造について、語基あるいは部分複合名詞間の係り受け尤度を合計し、全体の構造の尤度が最も高いものを採用する。

8 実験と考察

日本経済新聞 CD-ROM93 年版の中に出現する複合名詞を使用して解析する実験を行った。4 あるいは 5 の語基からなる出現頻度の高い複合名詞を機械的に抽出し、提案手法により解析を行った。部分複合名詞の分割と係り受け尤度に用いる共起情報は、年度の違う新聞 (日本経済新聞 CD-ROM94 年版) から 7.1 で述べた複合名詞列を抽出して使用した。部分複合名詞と認定する *C-value* の閾値 T_c は、複合名詞に含まれる語基および部分複合名詞 $w_i(1 \leq i \leq n)$ としたとき次式で決定した ($const > 1$)。

$$T_c = \frac{\sum_i^{n-1} C\text{-}value(w_i w_{i+1})}{n-1} * const \quad (3)$$

また格フレーム辞書および意味属性には日本語語彙大系 [6] のものを用いた。

比較のため、部分複合名詞の分割を行わず bigram による共起頻度のみを使用して解析する方法、最左導出によって決定する方法についても同様の実験を行った。(図 7)。

		提案手法	部分分割なし	最左導出
4 語基	正解数	50 (64 %)	34 (44%)	32 (41%)
	解析数	78	78	78
5 語基	正解数	43 (59 %)	29 (39%)	25 (34%)
	解析数	74	74	74

図 7: 解析結果

相互情報量のみを用いて解析を行った結果よりも 20 ポイント向上していることがわかる。この大きな要因の一つは相互情報量では欠落している頻度の絶対量を部分複合名詞に分割する際に考慮に入れている点であると考えられる。

部分分割による効果の例を図 8 示す。部分分割を使わない方法では、(二次, 戦略兵器) の相互情報量が高いために (第, 二次) と (戦略兵器) で先に構造をつくってしまい誤った構造をつくっている。しかし、実際にコーパス中の出現頻度は 8 回に過ぎず、関連度を過大に評価していると考えられる。一方、部分分割を使うとコーパス中の出現頻度と結合する語の多様性が考慮されるため、(戦略兵器, 削減) の方が強い結束性を持つと判断することができる。図 9 に解析結果例を示す。得られた構造では、部分複合名詞の単位で対応する英語訳に変換できることがわかる。

部分分割なし : (((第, 二次) 戦略兵器) , (削減, 条約))

部分分割あり : (@(第, 二次) , @(@(戦略兵器, 削減), 条約))

図 8: 部分分割による解析例

((@(外国, 為替), 変動),@(準備, 金))
(reserve) for (foreign exchange) (fluctuation)

(要約, @(貸借, @(対照, 表)))
(condensed) (balance sheet)

図 9: 解析結果例と英語訳

9 おわりに

コーパス中の共起情報を使い、始めに部分複合名詞に分割した上で、語基と部分複合名詞の構造を格フレーム辞書と共起情報を用いて決定する方法について検討した。解析段階で各語基の係り受け関係を同等に扱う方法と比べて、解析精度が良いだけでなく、日英翻訳の変換処理で扱い易い解析結果が得られることが確認された。

今後は、変換処理と組み合わせて複合名詞に適した翻訳方式を検討する予定である。

参考文献

[1] 西野哲朗, 藤崎哲之助. 漢字複合語の確率的構造解析. 情報処理学会論文誌, Vol. 29, No. 11, pp. 1034–1042, 1988.

[2] 小林義行, 山本修司, 徳永健伸, 田中穂積. 語の共起を用いた複合名詞の解析. 情報処理学会研究報告, Vol. 101-1, pp. 1–8, 1994.

[3] 久光徹. 文書走査を用いた複合名詞解析について. 情報処理学会研究報告, Vol. 112-2, pp. 7–14, 1996.

[4] 宮崎正弘, 池原悟, 横尾昭男. 複合語の構造化に基づく対訳辞書の単語結合型辞書引き. 情報処理学会論文誌, Vol. 34, No. 4, pp. 743–753, 1993.

[5] Katerina T. Frantzi, Sophia Ananiadou, 辻井潤一. 専門用語の自動抽出. 情報処理学会研究報告, Vol. 112-12, pp. 83–88, 1996.

[6] 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.