

係受け共起頻度を利用した複合名詞の解析*

4P-4

内山将夫

板橋秀一†

筑波大学‡

1 まえがき

複合名詞は名詞(素)の連なった表現であり、その解析は形態素解析の精度の向上などの観点から重要である([1][2])。また、すべての複合名詞を名詞として辞書に登録するのは現実的ではない。

本システムでは共起頻度を利用して名詞素間の係受け関係を解析する([3])。共起関係の情報を形態素解析に利用することにより、一度学習された複合名詞を正しく切り分けられることを示す。以前に係受け関係を生じた名詞素間には再び係受け関係が生じやすく、共起頻度が高い名詞素のほうに係ることがその逆の場合よりも多い、という考え方方が係受け解析の基本方針である。

2 用語

a) 名詞素 複合名詞中で、意味の最小単位をなす漢字列をいう。名詞素への切り分けはユーザーの判断に任せられる。

b) 係受け素性 係受けの性質が同じ名詞素を統括する概念、複合名詞内には現れない。ユーザーが設定する。

c) 関係及び属性 名詞素Aと名詞素Bのあいだに係受け関係cがあるとき、これをA-[c]-Bと表す。

1. AがBの内的属性を規定しているとき、cを属性関係ということにする。

2. AがBの外的関係を表すとき、cを論理的関係ということにする。

例えば、「国際貢献」の係受け関係は、論理的関係「に格」により表され、「金メダル」の係受け関係は、属性関係「材質」によって表される。

d) 共起関係 共起関係として、以下のものを考える。

1) 直接共起 A-[r]-B のとき、AとBは係受け関係rで直接共起したといふ。

2) 間接共起 A-[r1]-C、B-[r1]-C、A-[r2]-D のとき、BとDは係受け関係r2で間接共起したといふ。図(1)

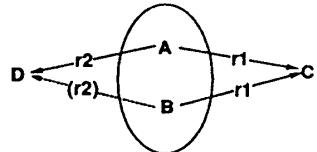


図1: 間接共起(r2)

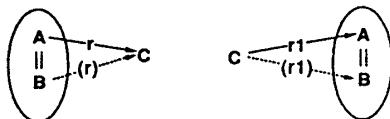


図2: 並列共起(r),(r1)

3) 並列共起 A-[と]-B、A-[r]-C のとき、BとCは係受け関係rで並列共起したといふ。また、A-[と]-B、C-[r1]-A のとき、CとBは係受け関係r1で並列共起したといふ。図(2)

4) 係受け素性を介した共起 AとBが同じ係受け素性を持ち、A-[r]-Cであるとき、BとCは係受け素性を介して係受け関係rで共起したといふ。また、C-[r1]-Aであるとき、CとBは係受け素性を介して係受け関係r1で共起したといふ

5) 属性関係による共起 A-[r]-B、C-[r]-D のとき (rは属性である)、AとD、CとBは、それぞれ属性rにより共起した、といふ。

これは、与える属性と規定される属性による名詞素間の意味的なつながりを示している。

6) 論理的関係による共起 係受け関係が論理的関係によるものということを除いて、属性関係による共起と同じである。

e) 共起頻度 名詞素間に生じた係受け関係の回数。

f) 係受け共起頻度 直接共起した名詞素と、その間に生じた係受けパターンに、共起頻度の情報を加えたもの。

名詞素は、辞書中に係受け共起頻度の情報を持つ。

*A Method of Compound Noun Analysis Utilizing Dependency Cooccurrence Frequency

†UTIYAMA Masao and ITAHASI Shuichi

‡University of Tukuba

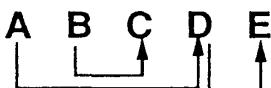
g) 不明語 複合名詞中のどの名詞素とも直接共起関係にないため、係受け関係が未知となっている名詞素。

3 解析

主な解析手順について説明する。

3.1 形態素解析

入力された平仮名文字列を名詞素列に切り分け、それぞれの名詞素列に優先度を付け加える。優先度とは、 $(0 <) \times (\text{名詞素間の直接共起の数} + 1) / \text{名詞素数}$ をいう。例えば、形態素解析の結果の名詞素列を ABCDE として、AD 間、BC 間、DE 間で直接共起関係が過去に生じたとする。このときの優先度の値は、 $(3 + 1) / 5 = 0.8$ である。



3.2 係受け解析

優先度が高い名詞素列から解析する。1、2、の順番で係受け解析が進行する。

1. 基本規則を満たす係り先候補に、直接共起に関する規則を適用して係り先を決定する。係り先が決定しない名詞素については、その決定を保留しておく。
2. 係り先の決定していない名詞素について、基本規則と不明語に関する規則を適用して係り先を決定する。

以下に、それぞれの規則を示す。

基本規則 全ての係受けに適用される規則である。

- 1 名詞素は自分より後ろの名詞素に必ず係る。
最後の名詞素は自分自身に係る。
- 2 係受け関係は交差しない。

直接共起に関する規則 直接共起を用いて係り先を決める場合の規則である。係受け関係は直接共起の情報から得られるものである。

1. 共起頻度が高いほど係受け関係が成立しやすい。
2. 共起頻度が同じ名詞素が複数個あった場合は、近い方の名詞素に係りやすい。

不明語に関する規則 直接共起により係り先が決定しなかった名詞素について、以下の共起関係を満たす名詞素をその係り先とする。

- 1) 間接共起
- 2) 並列共起
- 3) 係受け素性を介した共起
- 4) 属性関係による共起
- 5) 論理的関係による共起

共起関係による制約は、その番号順に処理が進行する。また係受け関係は共起関係から得られるものである。

4 考察

本システムでは、形態素解析に、通常の文における文節数最小法に直接共起の情報を取り入れたものを優先度として使用している。これにより、一度係受け関係が学習された複合名詞は、正しい切り分けである名詞素列の優先度が 1 (以上) になり、優先的に係受け解析をうける。不明語を含む場合は、複合名詞中の名詞素の数が多くなるほど、正しい切り分け列の優先度が高くなることが期待できる。(直接共起の数が多くなると考えられるから。)

共起関係は 2 名詞素の関係のみを表すため、複合名詞中に 3 名詞素以上の関係が成り立っているときには、係受け解析が失敗する場合がある。例えば、十月三日 (十は日と、月は日と、三は日と直接共起している) などの例では、係受け解析の結果として、

十 → 日、月 → 日、三 → 日

という誤った結果を出す。しかし、このような例はそれほど多くないと思われる (372 例中 6 例)。3 つ以上の構造を持つ係受け関係でも、2 名詞素間の関係によってうまく表される場合が多いのである。これは、複合名詞の構築に際して、なるべく単純な構造を使うことが意識されているからであろう。

5 むすび

形態素解析に係受け共起頻度の情報を利用することは適切といえる。係受け共起頻度のみで構文解析をすることは、通常の文に対しては単純に過ぎるが、複合名詞についてはかなり適切といえる。

係受け素性の生成と名詞素への割当を、直接共起の情報から自動生成することが、今後の課題である。

参考文献

1. 石崎雅人: 日本語複合名詞の解析、第 35 回情報処理学会全国大会論文集、1T-1(1987).
2. 石崎雅人: 2 名詞漢字複合名詞内の名詞の意味の多義解消アルゴリズム、情報処理学会論文誌、31、11、pp.1696-1699(1990).
3. 福田、板橋: 係り受け結合頻度を用いた複合名詞解析の一方法、情報処理学会第 42 回全国大会、1C-5、(1991).