決定リストを利用した形容動詞の修飾先の決定

美野 秀弥[†]　橋本 泰一[††]　徳永 健伸[††]　田中 穂積[††]

[†] 東京工業大学 工学部 情報工学科

[††] 東京工業大学 大学院 情報理工学研究科 計算工学専攻

{hide , taiichi , take , tanaka}@cl.cs.titech.ac.jp

1 はじめに

本研究では、「N_1 の A N_2」(N_i は名詞、A は形容動詞）という名詞句における形容動詞の修飾先を決定する手法を提案する。

1. 日本の豊かな自然
2. 緑の豊かな自然

例1では「豊かな」のは「自然」であるので、「豊かな」は名詞$_2$「自然」を修飾するのに対し、例2では「豊かな」のは「自然」ではなく、「緑」であるので、「豊かな」は名詞$_1$「緑」を修飾する。このように、「N_1 の A N_2」という名詞句において、形容動詞の修飾先が名詞$_1$であるか名詞$_2$であるか曖昧である。このような曖昧性を解決するには、構文的制約よりも意味的制約を重視する必要がある。

田中と荻野は、「N_1 の a/A N_2」(a は形容詞) という構文の形容詞/ 形容動詞の修飾先は、名詞$_1$と名詞$_2$の意味的依存関係と関連があると述べており、形容詞/形容動詞の修飾先を決定する「名詞関係依存の原則」を提案した [7]。「名詞関係依存の原則」は以下のとおりである。

1. 「N_1 の N_2」が成立すれば、形容詞/形容動詞は後ろの名詞$_2$を修飾する
2. 「N_2 の N_1」が成立すれば、形容詞/形容動詞は前の名詞$_1$を修飾する
3. 原則 1,2 が共に成立するならば、形容詞/形容動詞は名詞$_2$、名詞$_1$との間の結合可能性をさらに調べる

この原則は、形容詞/形容動詞の修飾先を決定するために、形容詞/形容動詞間の関係に着目するのではなく、前後の名詞間の関係のみに着目している点が興味深い。

橋本らは、名詞関係依存の原則に基づき、コーパスから得られる統計情報を用いて形容詞の修飾先を決定する手法を提案している [5]。「N_1 の N_2」と「N_2 の N_1」の共起事例のコーパスにおける出現頻度を利用し、共起頻度が大きい方を成立するとみなし、名詞関係依存の法則にしたがって形容詞の修飾先を決定している。この手法の正解率は約 90% であった。橋本らの手法では、名詞関係依存の原則 1,2 のみを適用しており、原則 3 については考慮していない。

菊池と伊藤は、約 4400 例文を人手で分析し、「N_1 の a/A N_2」という名詞句の形容詞/形容動詞の意味的依存関係を決定する 7 つの規則を提案している [4]。菊池と伊藤が作成した規則は、前後の名詞の特性で修飾先を決定する規則、形容詞/形容動詞と名詞の関係で修飾先を決定する規則、形容詞/形容動詞の特性で決定する規則の 3 つのカテゴリに分類されており、この規則によって、約 96% の正解率で形容詞/形容動詞の修飾先を決定することができたと報告している。しかし、人手で作成した規則では、多くの形容詞/形容動詞に対応することが困難となる。

白井らは、「N_1 の a N_2」という名詞句の形容詞の修飾先を決定するモデルを決定リスト [2] を用いてコーパスから自動的学習する手法を提案している [6]。白井らは単語間の意味的な制約を反映した決定リストを使用して、形容詞の修飾先を決定している。決定リストを用いることで、約 94% の正解率で「N_1 の a N_2」という句の形容詞の修飾先を決定することができたと報告している。

本研究では、「N_1 の A N_2」という名詞句において、白井らと同様に決定リストを用いて規則を学習し、データの過疎化に対応するために規則を抽象化する手法を提案する。また、形容動詞の修飾先と形容動詞がとりうる格との関係について考察する。形容動詞と修飾先の名詞句との結びつきと格挿入の間には密接な関係があると考えられる。

2 決定リストの学習

本節では、形容動詞の修飾先を決定する決定リストの学習アルゴリズムの詳細について述べる。用いるアルゴリズムは白井らが採用している手法とほぼ同じである [6]。しかし、白井らが考慮していない規則の抽象化の手法を提案する。

表 1 「社会経済情勢の急激な変化」から抽出できる規則

＜白井らの手法で抽出した規則＞				＜抽象化した規則＞			
抽出する規則		修飾先	タイプ	抽出する規則		修飾先	タイプ
A=急激 N_1=情勢 N_2=変化	→	N_2	an_1n_2	A=急激 N_1=113012 N_2=変化	→	N_2	aN_1n_2
A=急激 N_1=情勢	→	N_2	an_1	A=急激 N_1=113012	→	N_2	aN_1
A=急激 N_2=変化	→	N_2	an_2	N_1=113012 N_2=変化	→	N_2	N_1n_2
N_1=情勢 N_2=変化	→	N_2	n_1n_2	N_1=113012	→	N_2	N_1
N_1=情勢	→	N_2	n_1	A=急激 N_1=情勢 N_2=〈サ変名詞〉	→	N_2	an_1N_2
N_2=関係	→	N_2	n_2	A=急激 N_2=〈サ変名詞〉	→	N_2	aN_2
A=急激	→	N_2	a	N_1=情勢 N_2=〈サ変名詞〉	→	N_2	n_1N_2
				N_2=〈サ変名詞〉	→	N_2	N_2
				A=急激 N_1=113012 N_2=〈サ変名詞〉	→	N_2	aN_1N_2
				N_1=113012 N_2=〈サ変名詞〉	→	N_2	N_1N_2

2.1 規則の候補の抽出

品詞タグが付与されたコーパスから「N_1 の A N_2」というパターンを抽出し、それに正解を付与することで訓練データを作成する。但し、名詞句が複合名詞であった場合には、それらを 1 つとみなす。白井らの手法と同様に、名詞 $_1$、形容動詞、名詞 $_2$、その組合せによる規則を抽出する。

2.2 規則の抽象化

全ての単語を網羅する規則を抽出するためには、非常に多くのデータが必要となる。しかし、人手で正解を付与しているため、正解付き訓練データを劇的に増加させることは難しい。本研究では、この問題に対処するため、分類語彙表 [3] の意味クラスを利用し、単語を抽象化する。

また、菊池と伊藤は、特定の名詞が句に出現した場合、形容詞/形容動詞の修飾先が決定できると述べている。この分析をもとに決定リストの規則を抽象化する。抽象化する規則は以下のとおりである。

1. 人の属性を持つ名詞
2. サ変名詞
3. 数詞、単位などの属性を持つ名詞

単語の抽象化に関しては、分類語彙表の上位 6 桁の意味クラスを使用した。特殊名詞の抽象化に関しては、1 に関しては人であると思われるものを分類語彙表の上位 3 桁を基本にして人手により抽出し、2、3 に関しては RWC コーパスの品詞タグを利用した。デフォルト規則に関しては、形容動詞は名詞 $_2$ に修飾するという規則を適用した。

「社会経済情勢の急激な変化」という例文から抽出できる規則は表 1 のようになる。

2.3 決定リスト作成

2.3.1 尤度による順位付け

抽出した全ての規則に対して、それぞれ尤度を計算し、尤度の大きい規則を優先して順位付けを行なう。以下のように、規則 $r_i(C_i \rightarrow a)$ の尤度 $L(r_i)$ を算出する。

$$L(r_i) = \left| log_2 \left(\frac{P(a|C_i)}{P(\bar{a}|C_i)} \right) \right| \quad (1)$$

$P(a|C_i)$ は、条件 C_i を満たすときに形容動詞の修飾先が a(名詞 $_1$ または名詞 $_2$) になる確率を表す。その確率は以下のように定義する。

$$P(a|C_i) = \frac{O(a, C_i) + \alpha}{O(C_i) + \alpha} \quad (2)$$

$O(C_i)$ は、条件 C_i を満たす事例の総数であり、$O(a,C_i)$ は、条件 C_i を満たし、形容動詞の修飾先が a である事例の数を表している。α はスムージングのための値であり、本研究では 0.5 とした。

以下、尤度による順位付けを、DL_{log} と略記する。

2.3.2 規則のタイプによる順位付け

尤度による順位付けの場合、確かな規則であるのにも関わらず、データの頻度が少ないために適用されないことがある。一般性が高い規則 (タイプ a など) より特殊な規則 (タイプ an_1n_2 など) を先に適用する方が好ましいと考えられる。

そこで、人間が判断して規則の優先度を決定する。しかし、判断しがたい規則同士は、訓練データに適用したときの正解率を利用し規則の優先度を決定した。

以下、規則のタイプによる順位付けを DL_{type} と略記する。規則のタイプの優先順位は以下のように決定した。

an_1n_2 〉 aN_1n_2,an_1N_2 〉 aN_1N_2 〉 an_1,an_2,n_1n_2 〉 aN_1,aN_2,n_1N_2,N_1n_2 〉 a,n_1,n_2 〉 N_1N_2 〉 N_1,N_2,default

表 2 実験結果

	BaseLine	単語の抽象化なし		単語の抽象化あり	
		DL_{log}	DL_{type}	DL_{log}	DL_{type}
正解率	92.69% (241/260)	96.15% (250/260)	98.46% (256/260)	96.53% (251/260)	97.69% (254/260)
被覆率	100% (260/260)	63.46% (165/260)	96.54% (251/260)	80.77% (210/260)	96.54% (251/260)

3 評価実験

3.1 学習

毎日新聞の 1991 年から 1995 年までの新聞記事に対して、品詞タグを自動的に付与した RWC コーパス [1] を用いて本手法の評価実験を行なった。「N_1 の A N_2」というパターンにマッチする句を自動的に抽出し、1995 年 1 月から 5 月までのデータに関して、人手により形容動詞の修飾先を決定した。句として成立しないものと形容動詞の修飾先が一意に決まらないものに関しては除外した。形容動詞の修飾先が決定できた句の数は、2,075 である。この内、1,815 を訓練データとして、残りの 260 をテストデータとして使用した。

BaseLine はデフォルト規則のみを適用したものとする。そして、抽象化なしの場合とありの場合それぞれに対して DL_{log}、DL_{type} の手法で計 4 種類の実験を行なった。

3.2 実験結果と考察

実験結果を表 2 に示す。正解率はどの手法も BaseLine よりも大幅に正解率が向上している。したがって、本手法が有効的であると考えられる。

データの過疎化の問題に対処する方法として分類語彙表を使用して名詞の抽象化を行ない、規則を加えた。正解率は、尤度による順位付けの場合は抽象化することで増加し、規則のタイプによる順位付けの場合は抽象化することで減少した。これは、尤度による順位付けの場合はデフォルト規則の尤度以下の規則を適用しなかったのに対し、規則のタイプによる順位付けの場合は非常に尤度の低い規則が適用されることがあることが原因であった。規則の尤度とタイプの信頼度の両方を考慮する決定リストの適用順位の決定手法を考える必要がある。

また、尤度による順位付けの正解率が約 96% だったのに対し、規則のタイプによる順位付けの正解率は、約 98% と規則のタイプによる順位付けの方が有効であった。本研究では、人間が判断して、規則のタイプの適用順位を決定したが、さらに規則のタイプが増えた場合、人手により判断するのは困難である。自動的にタイプの適用順位を決定する手法を検討する必要がある。

3.3 格に関する考察

「N_1 の A N_2」という名詞句において N_1 と N_2 がどのような格をとりうるのか、そして、形容動詞の修飾先と格の関係について考察する。

「N_1 の A N_2」は主に以下の 7 つに分類できる。

1. 「名詞は名詞が形容動詞」に置き換えられる場合
   (a) ガ格の名詞が主格の役割の場合
   (b) ガ格の名詞が目的格の役割の場合
2. 「名詞は名詞が形容動詞」に置き換えられない場合
   (a) 「A N_2」に着目できる場合
      i. 形容動詞の意味が名詞と一部類似している場合
      ii. 形容動詞が名詞の表す対象自体ではなく対象の特徴的要素に係る場合
      iii. 形容動詞が名詞の対象のあり方と関係する場合
   (b) 「N_1」に着目できる場合
   (c) 「N_2」に着目できる場合

1-(a) → ガ格を持つ名詞句を修飾

例) 日米の親密な関係 → 日米は関係が親密だ

「日米は関係が親密である」と解釈でき、ガ格を持つ名詞句は主格の役割を持っている。このとき、「親密な」は、主格となる「関係」を修飾する。

1-(b) → ガ格を持つ名詞句を修飾

例) 外国語の苦手な私 → 私は外国語が苦手だ

「私は外国語を苦手としている」と解釈でき、ガ格を持つ名詞句は目的格の役割を持っている。このとき、「苦手な」は、目的格となる「外国語」を修飾する。

2-(a) → N_2 を修飾

神崎は「形容詞類 名詞類」の結びつき方にいくつかのタイプが見られると述べている [8]。以下にそれらのタイプを示す。

2-(a)-i

例) 彼の圧倒的な迫力

「迫力」の表す対象と「圧倒的な」の表す程度が一部重複しているので、語と語の結び付きが強く、「圧倒的な」は「迫力」を修飾すると解釈できる。しかし、重複しているために「迫力が圧倒的だ」と格を挿入すると、表現が不適切になる。

2-(a)-ii

例) 占星術の異常な愛好者

「異常な」は「愛好者」ではなく、愛好性の度合を修飾すると考えられる。そのため、「愛好者が異常だ」と格を挿入すると解釈にずれが生じる。

2-(a)-iii

例) 幹部間のひそかな会合

「ひそかな」は「会合」の様態、つまり置かれている状況という外的な要因によって関係を結んでおり、「ひそかな」は「会合」を修飾すると考えられる。しかし、「会合がひそかだ」と格を挿入すると表現が不適切になる。

2-(b) → N_2 を修飾

例) 千五百本の真っ赤なカーネーション

「千五百本」は「真っ赤だ」とは直接関わりを持てず、「カーネーション」を限定することしかできないために格を挿入することができないと考える。

菊池と伊藤は、N_1 が数詞、単位、まとめる語、時詞、位置の場合に着目しているが、これらの場合がこのパターンであると考えられる。

2-(c) → N_1 を修飾

例) 本の好きな理由

この句は、名詞$_2$「理由」が修飾節の内容に対して概念的に相対している。つまり、「本が好きだ」という出来事が生じた後、その結果として名詞$_2$が生じていることを表している。このような名詞$_2$は他には、「原因」、「結果」などがある。阿辺川は、このような連体修飾節を「相対概念型」と呼んでいる [9]。

4 おわりに

本研究では、「N_1 の A N_2」という文中の A の修飾先を決定する決定リストを学習する手法を提案した。人手により形容動詞の修飾先を付与した名詞句より、規則の候補を抽出した。そして、規則の信頼度を表す尤度により順位付けされた決定リストと正解率の高い規則のタイプによって順位付けされた決定リストを作成し、作成した決定リストを用いて形容動詞の修飾先を決定した。また、少ないデータで有効な規則を抽出するために、意味クラスや品詞による名詞の抽象化を行なった決定リストも作成手法を示した。そして、評価実験において約 98% の正解率で正しい形容動詞の修飾先を決定することができ、本手法が有効的であることを確認した。

参考文献

[1] Koiti Hashida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. TheRWC Text Databases. In Proceedings of the First International Conference on Language Resources and Evaluation, pp. 457-462,1998.

[2] Ronald L. Rivest Learning decision lists. Machine Learning, Vol.2,pp.229-246,1987.

[3] 中野洋. 「分類語彙表」形式による語彙分類表 (増補版) 第 1 分冊<本表>, 第 2 分冊<索引>. 国立国語研究所言語体系研究部,1996.

[4] 菊地浩三, 伊藤幸宏. 連体形イ・ナ形容詞に先行する格助詞句の係りに関するルールの抽出. 言語処理,Vol.6,No.3,pp.75-99,1999.

[5] 橋本泰一, 白井清昭, 徳永健伸, 田中穂積. 統計的手法に基づく形容詞または形容動詞の修飾先の決定. 情報処理研究報告 (2000-NL-138),Vol.2000,Vol.65,pp.87-94,2000.

[6] 白井清昭, 橋本泰一, 西館耕介, 徳永健伸, 田中穂積. 決定リストを用いた形容詞の修飾先の決定. 言語処理学会第 7 回年次大会発表論文集, pp.253-256,2001.

[7] 田中穂積, 荻野孝野. 形容詞もしくは形容動詞の修飾先の名詞を決める原則について. 計量国語学,Vol.12,No.5,pp.191-203,1980.

[8] 神崎享子. 連体修飾関係を結ぶ形容詞類と名詞. 計量国語学, Vol21.No.2,pp.53-67,1997.

[9] 阿辺川武. 統計情報を利用した日本語連体修飾節の解析. 東京工業大学大学院情報理工学研究科, 修士論文,2001.