

決定リストにおける規則の適用順序に関する考察

八木 豊 橋本 泰一 美野 秀弥 徳永 健伸 田中 穂積

東京工業大学大学院 情報理工学研究科
〒 152-8552 東京都目黒区大岡山 2-12-1

E-mail: {yutaka,taiichi,hide,take,tanaka}@cl.cs.titech.ac.jp

決定リストは、自然言語処理における様々な問題に応用され、その有効性が示されている。決定リストを用いるうえでの重要な問題は、どのような基準に基づいて規則の優先順位を与えるかということである。本論文では、判定の根拠となる証拠によって規則をタイプごとに分類し、規則のタイプの定性的な評価に基づいて優先順位を与える手法を提案する。提案手法を評価するために、特定の構文における形容詞/形容動詞の修飾先の決定と、語の多義性解消のコンテスト SENSEVAL-2 で行われた日本語辞書タスクという2つのタスクで実験を行い、従来の手法と比較した。その結果、いずれのタスクにおいても提案手法と対数尤度比を用いる従来の手法に有意な差はみられなかった。

On rule ordering in decision lists

Yutaka YAGI, Taiichi HASHIMOTO, Hideya MINO, Takenobu TOKUNAGA, and
Hozumi TANAKA

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
Ō-okayama Meguro-ku Tokyo 152-8552 JAPAN

E-mail: {yutaka,taiichi,hide,take,tanaka}@cl.cs.titech.ac.jp

Decision lists have been used in various natural language processing tasks. The order of decision rules are usually decided based on their log likelihood calculated from training data. This paper proposes a new rule ordering method in which the order of rules is decided by considering information type used by rules as well. The proposed method was evaluated in two tasks: (1) dependency analysis of adjectives in Japanese noun phrases, and (2) Japanese dictionary-based task in SENSEVAL-2. The experiments showed the proposed method did not show significant improvement in these two tasks.

1 はじめに

決定リスト [3] は、クラス分類を行う機械学習手法の 1 つである。自然言語処理における問題の多くはクラス分類の問題として捉えることができることから、アクセント記号復元 [9]、単語のわかち書き [5]、形容詞の修飾先の決定 [6]、語の多義性解消 [8] など、様々な問題に決定リストを応用した研究が報告され、その有効性が示されている。

決定リストを作成するうえでの重要な問題は、どのような信頼度を考慮して規則の優先順位を与えるかである。最尤推定によって信頼度を推定する従来の手法では、事例の数が少ないときにその値が不正確になることから、鶴岡らは、ベイズ推定によって信頼度の推定を行い決定リストの性能を向上させている [7]。

新納は、同音異義語の判別において、複合語の一部であることから作成される証拠がその他の証拠よりも信頼できることを考慮し、そのような証拠に基づく規則に重みを与える手法を提案している [4]。

白井らは、「 N_1 の *Adj* N_2 」(ただし、 N_1, N_2 は名詞、*Adj* は形容詞とする) という特定の構文における形容詞の修飾先が、 N_1 と N_2 の意味的依存関係と関連することを考慮し、 N_1 と N_2 に基づく規則を優先的に適用する手法を提案している [6]。

本論文では、上記の複合語や名詞間の意味的依存関係のように、対象とする分類問題に対して、ある特定のタイプの証拠に基づいた規則が有効に働くことを仮定し、規則のタイプの定性的な評価に基づいて優先順位を与える手法を提案する。また、提案手法を評価するために、特定の構文における形容詞/形容動詞の修飾先の決定と、語の多義性解消のコンテスト SENSEVAL-2[2] で行われた日本語辞書タスクという 2 つのタスクに対して実験を行った。

本論文の構成を以下に示す。2 節では、決定リストの各規則に優先順位を与える従来の手法と、本論文で提案する手法について述べる。次に 3 節で、提案手法の評価に用いた 2 つのタスクの詳細を説明し、4 節で、それぞれのタスクで実際に実験を行った結果を示す。最後に 5 節で、本論文のまとめを行う。

2 決定リスト

決定リストは、式 1 で表されるような if-then 形式の規則の集合である。

$$r_i : E_i \rightarrow d_i \quad (1)$$

ここで、規則 r_i は、ある事例が証拠 E_i を満たすときに、分類対象に対して d_i という判定をすることを表す。1 つ 1 つの規則は、なんらかの信頼度に基づいて優先順位が与えられる。決定リストによる評価では、新たな事例ごとに、優先順位の上位の規則から証拠を調べてゆき、与えられた事例が初めて証拠を満たした規則を適用する。したがって、どのような信頼度を考慮して規則に優先順位を与えるかが重要になる。

2.1 尤度に基づく順位付け

従来の手法では、規則 r_i の信頼度に式 2 で定義される対数尤度比、あるいは式 3 で定義される条件付き確率を用いる。

$$(\text{信頼度}) = \log \frac{P(d_i | E_i)}{P(\bar{d}_i | E_i)} \quad (2)$$

$$(\text{信頼度}) = P(d_i | E_i) \quad (3)$$

ここで、 $P(d_i | E_i)$ は証拠 E_i が成立するときに判定 d_i を行う確率であり、 $P(\bar{d}_i | E_i)$ はその排反事象である。これは、式 4 で推定する。

$$P(d_i | E_i) = \frac{O(E_i, d_i) + \alpha}{O(E_i) + \alpha} \quad (4)$$

ここで、 $O(E_i)$ は証拠 E_i が成立する事例の出現頻度であり、 $O(E_i, d_i)$ は証拠 E_i が成立し、かつ判定が d_i である事例の出現頻度である。また、 α はスムージングのためのパラメタである。今回の実験では $\alpha = 0.5$ とした。

さらに、式 5 で表されるデフォルト規則を導入する。

$$\text{true} \rightarrow \text{most_frequent_decision} \quad (5)$$

これは、証拠を考慮せずに無条件で適用され、分類対象に対して最も頻度の高かった判定を行う規則である。デフォルト規則よりも尤度が低い規則が適用されることはないので、そのような規則は決定リストから除去する。

2.2 規則のタイプに基づく優先付け

本論文で提案する手法では、ある特定のタイプの証拠に基づいた規則が有効に働くことを仮定する。そこで、判定の根拠となる証拠によって規則をタイプごとに分類し、規則のタイプによって優先順位を与える。規則のタイプによる順序は、式6で定義するタイプの正解率 $prec(T_i)$ によって決定する。

$$prec(T_i) = macroaverage_{T_i \ni r_i} prec(r_i) \quad (6)$$

ここで、 $prec(r_i)$ は規則 r_i の正解率で、規則 r_i だけを用いて訓練データを分類したときに正しく分類できた事例の割合を表す。タイプの正解率 $prec(T_i)$ は、タイプ T_i に含まれるすべての規則に対して、規則の正解率を求めて、その値のマクロアベレージをとったものである。したがって、 $prec(T_i)$ の値が大きいほど、そのタイプの信頼度が高いことになる。

提案手法では、すべてのタイプの正解率を求めて、その値が大きい順に規則を並べる。ただし、同じタイプの規則の順序は、式2の対数尤度比あるいは式3の条件付き確率によって決定する。4節で示す実験では、対数尤度比を用いた。また、尤度に基づく順位付けで行ったデフォルト規則より尤度の低い規則の除去は行わない。

3 評価用タスク

本論文で提案する手法の評価を行うためのタスクとして、以下の2つのタスクを設定した。1つは、ある特定の構文において形容詞または形容動詞の修飾先を決定するタスク [6]、もう1つは、語の多義性解消のコンテストである SENSEVAL-2 で行われた日本語辞書タスクである [2]。本節では、それぞれのタスクの詳細を説明するとともに、各タスクで使用するデータ、分類に用いる証拠について述べる。

3.1 形容詞/形容動詞の修飾先の決定

3.1.1 タスクの概要

「 N_1 の Adj N_2 」および「 N_1 の $AdjV$ N_2 」(ただし、 N_1, N_2 は名詞、 Adj は形容詞、 $AdjV$ は形容動詞とする) という構文において Adj または $AdjV$ の修飾先を決定するタスクである。例文を以下に示す。

1. 信濃の美しい川
2. 水面の美しい川
3. 意識の急激な変化

4. あいさつの苦手な私

例文1では、形容詞「美しい」は後ろの名詞「川」を修飾し、例文2では、形容詞「美しい」は前の名詞「水面」を修飾する。同様に、例文3では、形容動詞「急激な」が後ろの名詞「変化」を修飾し、例文4では、形容動詞「苦手な」が前の名詞「あいさつ」を修飾している。このように、上記の構文における Adj や $AdjV$ の修飾先は曖昧である。 Adj や $AdjV$ の修飾先を正確に決定しなければならない場面は多い。例えば、機械翻訳で例文1と2の訳し分け、例文3と4の訳し分けをすることを考えれば、 Adj や $AdjV$ の修飾先を決定することの意義は明らかであろう。

本論文では、「 N_1 の Adj N_2 」という構文内の Adj の修飾先を決定するタスクと、「 N_1 の $AdjV$ N_2 」という構文内の $AdjV$ の修飾先を決定するタスクに分けて決定リストを学習する。

3.1.2 データ

このタスクにおける訓練データは「 N_1 の Adj N_2 」または「 N_1 の $AdjV$ N_2 」という構文の集合である。このうち、 Adj については、[6] で白井らが用いている正解が付与されたデータを利用する。ただし、まったく同じ事例が複数含まれているということがあったため、それらを取り除く作業を行った。

$AdjV$ に対する訓練データは、上記の構文に対応する文を品詞のパターンマッチによってコーパスから自動的に抽出し、人手で正しい修飾先を与えて新たに作成した。このとき、連続して現れている名詞は1つの複合名詞とみなし N_1 または N_2 として抽出した。コーパスには、RWC コーパスを用いた [1]。これは、毎日新聞 1991年から1995年の新聞記事に対して品詞タグを自動的に付与したコーパスであるが、用いたのはその一部分だけである。

用意した訓練データの数は、 Adj に対する構文が 3,744 文、 $AdjV$ に対する構文が 1,912 文である。評価データを別に用意することはせず、訓練データを n 分割して n -fold crossvalidation による評価を行う。

3.1.3 分類に用いる証拠

Adj の修飾先を決定するための情報として6種類の証拠を用いる。それに基づいた規則のテンプレートを表1に示す。このテンプレートに従って、例えば、「話題:満載の若い挑戦:者」という事例からは

表 1: 規則のテンプレート (*Adj*)

規則 ($E_i \rightarrow d_i$)	タイプ
$Adj = x \rightarrow \{N_1, N_2\}$	a
$N_1 = x \rightarrow \{N_1, N_2\}$	n1
$N_2 = x \rightarrow \{N_1, N_2\}$	n2
$Adj = x \ \& \ N_1 = y \rightarrow \{N_1, N_2\}$	a+n1
$Adj = x \ \& \ N_2 = y \rightarrow \{N_1, N_2\}$	a+n2
$N_1 = x \ \& \ N_2 = y \rightarrow \{N_1, N_2\}$	n1+n2

以下に示す 6 つの規則を生成する .

- $Adj = \text{若い} \rightarrow N_2$
- $N_1 = \text{満載} \rightarrow N_2$
- $N_2 = \text{者} \rightarrow N_2$
- $Adj = \text{若い} \ \& \ N_1 = \text{満載} \rightarrow N_2$
- $Adj = \text{若い} \ \& \ N_2 = \text{者} \rightarrow N_2$
- $N_1 = \text{満載} \ \& \ N_2 = \text{者} \rightarrow N_2$

このとき, N_1 や N_2 が「話題:満載」や「挑戦:者」のような複合名詞の場合には, 訓練データ全体から生成される規則の種類が増えるのを抑えるために, 複合名詞全体ではなく一番最後に現れる名詞のみを参照する .

AdjV についても *Adj* の部分を置き換えるだけの同様な証拠を用いた .

3.2 SENSEVAL-2 日本語辞書タスク

3.2.1 タスクの概要

SENSEVAL-2 は, 2001 年 7 月に行われた多義性解消のコンテストである . 日本語辞書タスクはそのとき行われたタスクの一つで, 単語の語義を岩波国語辞典によって定義し, テキスト中のあらかじめ決められた対象単語の語義を決定するタスクである . 対象単語は, 名詞 50 個, 動詞 50 個の合計 100 種類の単語となっている . 本論文では, 対象単語ごとに 1 つの決定リストを学習する .

3.2.2 データ

訓練データは, テキスト中の一部の単語に対して, 人手で語義が与えられたコーパスである . コーパスには, 語義情報の他に形態素情報 (わかち書き, 品詞, 読み, 基本形), UDC コードが与えられている . UDC コードとは, 国際十進分類法 (Universal Decimal Classification)[10] によるコード番号のことで, テキストを分類するための情報である . 実験では, さらに JUMAN と KNP を用いて構文情報

を与えた .

訓練データの数は, 1994 年の毎日新聞の 3000 記事, 含まれる形態素は約 900,000 である . 評価データは, 訓練データとは別の新聞記事から用意されていて, 1 つの対象単語につき 100 事例, 合計 10,000 個の事例からなる . 訓練データ同様, 形態素情報, UDC コードが与えられており, さらに JUMAN と KNP を用いて構文情報を与えた .

3.2.3 分類に用いる証拠

対象単語の語義を決定するための情報として 14 種類 8 タイプの証拠を用いる . それに基づいた規則のテンプレートを表 2 に示す . 表中の w_i は,

表 2: 規則のテンプレート (*S2J*)

規則 ($E_i \rightarrow d_i$)	タイプ
$w_{-1} = x \rightarrow s_i$	$W_{adjacent}$
$w_{+1} = x \rightarrow s_i$	
$p_{-1} = x \rightarrow s_i$	$P_{adjacent}$
$p_{+1} = x \rightarrow s_i$	
$w_{-2} = x \ \& \ w_{-1} = y \rightarrow s_i$	W_{pair}
$w_{-1} = x \ \& \ w_{+1} = y \rightarrow s_i$	
$w_{+1} = x \ \& \ w_{+2} = y \rightarrow s_i$	
$p_{-2} = x \ \& \ p_{-1} = y \rightarrow s_i$	P_{pair}
$p_{-1} = x \ \& \ p_{+1} = y \rightarrow s_i$	
$p_{+1} = x \ \& \ p_{+2} = y \rightarrow s_i$	
$w_{k, -4 \leq k \leq 4} = x \rightarrow s_i$	W_{window}
$VERB = x \rightarrow s_i$	verb
$OBJ = x \rightarrow s_i$	obj
$UDC = x \rightarrow s_i$	udc

多義性解消の対象となっている単語から i の位置にある単語を表す . ただし, 単語が活用語の場合には, その基本形を参照する . 同様に p_i は, i の位置にある品詞を表す .

VERB は, 対象単語が名詞の場合のみ用いられる証拠で, 対象単語が係っている動詞を表す . 一方, *OBJ* は, 対象単語が動詞の場合のみ用いられる証拠で, 対象単語の目的語を表す . 目的語が複合名詞の場合には, 複合名詞全体ではなく一番最後に現れる名詞のみを参照する .

UDC は, 対象単語を含むテキストの UDC コードを表す . ただし, UDC コードは主標数, 固有補助標数, 共通補助標数の組み合わせにより複雑な表現をされているため, 主標数のみを取り出して参照

する。

4 評価実験

4.1 形容詞/形容動詞の修飾先の決定

4.1.1 学習

このタスクの評価は 10-fold crossvalidation によって行う。したがって、RWC コーパスから取り出した「 N_1 の Adj N_2 」という構文 3,744 個、「 N_1 の Adv N_2 」という構文 1,912 個をそれぞれ 10 分割して、1 つを評価データに、残りを訓練データにして決定リストを学習する作業を 10 回行った。このとき、対数尤度比に基づいて優先順位を与えた決定リスト dl-L と、規則のタイプに基づいて優先順位を与えた決定リスト dl-T の 2 種類を学習した。ただし、dl-T については 2.2 項で示した手法であらかじめ予備実験を行い、その結果 (表 3,4) から規則のタイプによって以下に示す順序を与えた。

Adj : $n1+n2 > a+n1 > a+n2 > n1 > n2 > a$
 Adv : $a+n1 > a+n2 > n1+n2 > n1 > a > n2$

表 3: 予備実験の結果 (Adj)

規則のタイプ	正解率
a	84.4%
n1	97.5%
n2	92.0%
a+n1	99.9%
a+n2	99.8%
n1+n2	100%

表 4: 予備実験の結果 (Adv)

規則のタイプ	正解率
a	98.3%
n1	98.8%
n2	98.3%
a+n1	100%
a+n2	100%
n1+n2	100%

4.1.2 実験結果

表 5,6 は、10-fold crossvalidation によって得られた正解率と規則適用率である。正解率は修飾先を正しく判定できた割合、規則適用率は判定の正誤

表 5: 実験結果 (Adj)

	BL	DL-L	DL-T
正解率	52.6%	92.3%	91.4%
規則適用率	0.0%	98.9%	99.4%

表 6: 実験結果 (Adv)

	BL	DL-L	DL-T
正解率	89.0%	95.8%	95.0%
規則適用率	0.0%	61.9%	96.5%

に関わらずデフォルト規則以外の規則によって修飾先を判定できた割合である。BL(ベースライン)は、 Adj または Adv の修飾先を常に N_2 にする手法を表す。 Adj では、dl-L, dl-T とともに正解率が約 92%となり、ベースラインを大きく上回ったが、両者に有意な差はみられなかった。規則適用率は、いずれも約 99%であった。一方、 Adv は、 Adj と比較してベースラインがかなり高いが、dl-L, dl-T とともに正解率が約 95%となり、ベースラインを上回った。しかし、こちらも両者に有意な差はみられなかった。規則適用率では、dl-L と dl-T の間に大きな差があった。これは、ベースラインの正解率が高いためにデフォルト規則の対数尤度比が大きくなり、dl-L では多くの規則が除去されるのに対し、dl-T ではデフォルト規則による除去を行わないことによる。

4.2 SENSEVAL-2 日本語辞書タスク

4.2.1 学習

配布された訓練データ 3,000 記事と岩波国語辞典の語釈文に含まれる例文から、それぞれの対象単語に対する決定リスト dl-L と dl-T を学習した。dl-T については、予備実験の結果 (表 7) から規則のタイプによって以下に示す順序を与えた。

$verb, obj > w_{pair} > w_{adjacent} > p_{pair} >$
 $w_{window} > p_{adjacent} > udc$

これは、すべての対象単語に共通である。

4.2.2 実験結果

表 8 は、学習した決定リストを評価データに適用したときの正解率と規則適用率である。BL は、対象単語ごとに最も頻度の高い語義を選択する手法を表す。dl-L, dl-T とともに正解率は約 78%となり、ベースラインを上回ったが、両者に有意な差はみら

表 7: 予備実験の結果 (S2J)

規則のタイプ	正解率
<i>wadjacent</i>	83.8%
<i>Padjacent</i>	74.1%
<i>wpair</i>	84.7%
<i>Ppair</i>	79.5%
<i>wwindow</i>	77.7%
<i>verb</i>	87.4%
<i>obj</i>	88.9%
<i>udc</i>	72.8%

表 8: 実験結果 (S2J)

	BL	DL-L	DL-T
正解率	72.6%	77.91%	77.26%
規則適用率	0.0%	98.87%	100%

れなかった。規則適用率は、dl-T のほうで 100% になり、すべての判定をデフォルト規則以外の規則によって行うことができた。

5 おわりに

本論文では、判定の根拠となる証拠によって規則をタイプごとに分類し、規則のタイプの定性的な評価に基づいて優先順位を与える手法を提案した。特定の構文における形容詞/形容動詞の修飾先の決定と、語の多義性解消のコンテスト SENSEVAL-2 で行われた日本語辞書タスクという 2 つのタスクに対する実験では、提案手法と対数尤度比を用いる従来の手法に有意な差はなかった。

参考文献

- [1] Koiti Hasida, et al. The RWC text database. In *Proceedings of the LREC*, pp. 457–462, 1998.
- [2] 黒橋禎夫, 白井清昭. SENSEVAL-2 日本語辞書タスク. 電子情報通信学会技術研究報告 NLC2001-36 ~ 48, Vol. 101, No. 351, pp. 1–8, 2001.
- [3] Ronald L. Rivest. Learning decision lists. *Machine Learning*, Vol. 2, pp. 229–246, 1987.
- [4] 新納浩幸. 複合語からの証拠に重みをつけた決定リストによる同音異義語判別. 情報処理学会

論文誌, Vol. 39, No. 12, pp. 3200–3206, 1998.

- [5] 新納浩幸. 決定リストを弱学習器としたアダプティブによる日本語単語分割. 自然言語処理, Vol. 8, No. 2, pp. 3–18, 2001.
- [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] 鶴岡慶雅, 近山隆. ベイズ推定による決定リストのルール信頼度推定法. 情報処理学会研究報告 2001-NL-144, Vol. 2001, No. 69, pp. 91–97, 2001.
- [8] 八木豊, 野呂智哉, 白井清昭, 徳永健伸, 田中穂積. 決定リストを用いた語義曖昧性解消. 電子情報通信学会技術研究報告 NLC2001-36 ~ 48, Vol. 101, No. 351, pp. 47–52, 2001.
- [9] David Yarowsky. DECISION LISTS FOR LEXICAL AMBIGUITY RESOLUTION: Application to Accent Restoration in Spanish and French. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 88–95, 1994.
- [10] 情報科学技術協会. 国際十進分類法. 丸善株式会社, 第 3 版, 1994.