3ZE-08 帰納論理プログラミングを用いた順位づけ規則の帰納

中野 智文　　犬塚 信博　　伊藤 英則
名古屋工業大学

1 はじめに

本論文では帰納論理プログラミング (Inductive Logic Programming; 以下 ILP) を用いて順位づけのための知識を得る手法を提案する．順位事例を全順序関係とみてそこから比較関係の事例を得，その定義を帰納する．また帰納された比較関係を未知の対象に適用し順位を与える方法を検討する．この方法の有効性を確認するために英文品詞タグ付け問題へ応用し実験を行なう．

2 英文品詞タグ付け問題

英文品詞タグ付け問題は，英文中の個々の単語に品詞をつける問題である．例えば，“This is a pen.” という英文では，This が代名詞，is が動詞，という様に品詞が付けられる．このような品詞の情報より文法を解釈し，その英文のもつ意味の理解や翻訳などの自然言語の分野に役立てるのが英文品詞タグ付け問題の目的である．

タグづけ規則の学習方法の有効性を確認するために，全ての単語に品詞タグをつけられた英文（コーパス）を用いることができる．これを訓練用とテスト用にわけ，訓練用コーパスから事例として知識を獲得し，その知識を利用してテスト用コーパスの英文に対してタグづけを行ない，その正解率を確かめる．以下ではコーパス中のあるタグつき英文を，$s = [w_1/t_1, w_2/t_2, ..., w_n/t_n]$ と表す．ただし w は単語を，t は品詞タグである．

最も単純な方法は，頻度つき品詞辞書を用いる方法である．品詞辞書とは，個々の単語がとり得る品詞を単語ごとに収めた辞書のことであり，訓練用コーパスから作成できる．この品詞辞書を使うことで未知の英文の各単語に対して，それがとり得る品詞（品詞候補）で最も高い頻度で使われた品詞をそのタグとしてつけることができる．

3 Cussens のタグつけ手法

Cussens は，頻度つき品詞辞書が許す品詞候補を更に限定する知識を帰納的に獲得する方法を提案した [1]．つまり，予め品詞を限定する知識を帰納しておき，これを使って品詞辞書からの品詞の候補を限定し，次に頻度より最終的に決定する．

Cussens は候補限定のための知識の帰納に ILP を用いた．ILP は，機械学習の 1 つで，論理プログラムによって書かれた背景知識と，目標とする述語の正事例と負事例から，その述語の論理プログラムを帰納する学習手法である．背景知識に論理プログラムでかかれた文法的知識を与えることができ，言語のような構造的知識が必要な学習に向いているのが ILP を用いる利点である．

3.1 限定知識の帰納

Cussens の手法では，品詞候補の絞り込みのため，「この品詞タグを候補から除け」という意味の述語「rmv(remove)」を帰納する．文脈情報 c_1, c_2 と品詞タグ t に対し，述語 $\mathrm{rmv}(c_1, c_2, t)$ はその品詞タグ t が最終候補から除かれるべきなら真をそうでなければ偽を返す．ここでタグつき英文 $s = [w_1/t_1, ..., w_{i-1}/t_{i-1}, w_i/t_i, w_{i+1}/t_{i+1}, ..., w_n/t_n]$ に対し，タグを付けたい単語 w_i としたとき，c_1,c_2 は，文脈情報であり，各々 $[t_1, t_2, ..., t_{i-1}]$，$[t_{i+1}, ..., t_n]$ である．

rmv の定義を帰納するためにその事例を訓練コーパスから得る必要がある．

$\mathrm{rmv}([t_1, t_2, ..., t_{i-1}], [t_{i+1}, ..., t_n], t_i)$ を負事例

$\mathrm{rmv}([t_1, t_2, ..., t_{i-1}], [t_{i+1}, ..., t_n], t_c)$ を正事例

として得る．ここで，$t_c \in T - \{t_i\}$，T は品詞辞書が許す w_i の品詞タグである．つまり，真のタグ t_i を除くことは間違いであり，すなわち負事例である，逆にその他のタグ t_c は除くべきであり，すなわち正事例である．

Cussens はこれらの事例から述語 rmv の定義を ILP システムにより帰納できることを示した．

3.2 知識の適用と問題点

帰納された知識の適用を述べる．英文中の未知の品詞の単語に対し，品詞辞書を使って品詞の候補を求める．次にその候補中の個々の品詞に対して，帰納された述語 rmv が真を返すならば，その品詞を候補から取り除く．最後に，残った候補のうち頻度が最も高いものを解とする．

この手法では候補が 1 つに絞られにくいという問題点がある．この手法では解が 1 つであることを学習及び適用段階で考慮されていないためにこのような問題が生ずると考えられる．

もちろん，この問題の場合複数の候補があっても，頻度により決定する方法ため解決されている．しかし，知識によってより絞り込むことができればより多く正解を得ることができると考えられる．

4 順位づけ問題への変換

タグづけ問題は複数の品詞の候補から 1 つを選択する問題であった．我々はこの問題を一種の順位問題と考える．すなわち，これは候補集合に順位をつけ，1 位

Inducing ranking relation using ILP
Tomofumi Nakano, Nobuhiro Inuzuka and Hidenori Itoh
Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

を選ぶという順位問題に置き換えることができる．

一般に順位問題を，順位づけられる候補の集合（この問題では品詞の候補）$C = \{c_1, c_2, ..., c_n\}$ に対し，その解として C の要素を並べた列（順位列）を与える問題と定義する．我々はこれまでに状態空間を探索する問題において，探索すべき状態を順序付ける規則を帰納する手法を提案してきた [2]．

提案する手法は，集合 C の要素を比較する 2 項関係 cmp の定義を既知の順位事例から帰納し，これを用いて C を順序づける．一般に帰納される cmp の定義は線形順序ではないため，線形順序を推測する方法（順位づけ手法）を合わせて与える必要がある．

4.1 比較知識の帰納段階

Cussens の手法では各タグに対して「それを除け」という述語を帰納していたが，本手法では 2 つのタグのうち「こちらを選択すべき」という比較述語 cmp を帰納する．

$\mathrm{cmp}([t_1, t_2, ..., t_{i-1}], [t_{i+1}, ..., t_n], t_i, t_c)$ は正事例
$\mathrm{cmp}([t_1, t_2, ..., t_{i-1}], [t_{i+1}, ..., t_n], t_c, t_i)$ は負事例

第 1，2 引数は Cussens の手法と同じく，文中のその単語の前後の単語の品詞タグの情報が入る（以後これらを表記上省略する）．第 3，4 引数は品詞タグであり，ここここそが比較の関係で，第 3 引数のタグは第 4 引数のタグより選択するのにふさわしいという意味の述語である．なぜなら真のタグ t_i のほうが，他のタグ t_c より選択すべきであるからである．

これらの事例から述語 cmp を ILP システムにより帰納する．

4.2 帰納された比較知識による順位の決定

Cussens の手法では帰納した述語により候補を限定したが，提案手法では帰納した比較述語により各品詞候補間の関係を求め，その関係を元に各候補に順位をつける．

ここで，品詞辞書によって限定された品詞の候補の集合を $T = \{t_1, t_2, ..., t_m\}$ としたとき，帰納された比較述語 cmp を使う順位づけ手法はいくつかある．本論文では以下の 3 つを検討する．

(1) ラウンドロビントーナメント（リーグ戦方式）

$$t_i \geq t_j \text{ if } |\{t|\mathrm{cmp}(t_i, t) = true\}| \geq |\{t|\mathrm{cmp}(t_j, t) = true\}|$$

(2) ラウンドロビントーナメント改

$$t_i \geq t_j \text{ if } |\{t|\mathrm{cmp}(t_i, t) = true\}| - |\{t|\mathrm{cmp}(t, t_i) = true\}| \geq |\{t|\mathrm{cmp}(t_j, t) = true\}| - |\{t|\mathrm{cmp}(t, t_j) = true\}|$$

(3) 最小関係誤差

T の上の可能な線形順序 R のうち，cmp との誤差 Δ が最小なものに従う．

$$\Delta = |\{(t_1, t_2) \in T^2 | R(t_1, t_2) \neq \mathrm{cmp}(t_1, t_2)\}|$$

いずれの方法でも同立 1 位となることがある．その場合は複数の候補として残し，頻度に従う．

5 実験

実験では，ウォールストリートジャーナル約 300 万単語分のコーパスより約 200 万単語を訓練用コーパスとする．残りの 100 万単語をテスト用コーパスとする．訓練用コーパスで辞書を作成，知識を帰納し，これを用いてテストコーパスの英文をタグづけし正解率を確認する．

実験環境を Cussens と同じにするため，背景知識等は文献 [1] のものを取得，利用した．また知識を帰納するための ILP システムも [1] と同じ Progol を利用した．実験の対象はテストコーパスの英文中，品詞が一意に決定できる 75 万単語を除いた 25 万単語である．それらに対しタグづけを行なった．

結果を表 1 に示す．「正解」は帰納された知識と頻度による最終的なタグづけが正解だった単語の数を示す．「最終候補」は頻度による最終決定を行なう前の候補の状態を表し，「1」，「2 以上」は候補が 1 つまたは，2 以上でありかつそのなかには正解が含まれる単語の数，「間違い」はその候補に正解が含まれていない単語の数を示す．「最終候補」は，「1」が多くなればなるほど，「間違い」が小さくなればなるほど正しく 1 つの候補に絞り込む方法と考えられる．

表 1: 実験結果

手法	正解	最終候補 1	最終候補 2 以上	最終候補 間違い	全体
Cussens	217,018	134,475	110,215	12,071	256,761
提案 (1)	220,208	160,482	82,394	13,885	256,761
提案 (2)	220,212	161,852	80,432	14,477	256,761
提案 (3)	219,661	153,913	88,729	14,119	256,761

「提案」の (1)〜(3) は 4.2 の順位づけ手法 (1)〜(3) に対応．

Cussens の手法と比べ提案手法の正解単語数が増えていることが確認できる．目的であった最終候補の絞り込みも，「間違い」が増えているがそれ以上に「1」も増えており，正解数向上に貢献したと考えられる．

6 おわりに

選択問題を順位問題として扱う帰納論理プログラミングを用いた手法を提案した．また英文タグづけ問題の実験において，背景知識，事例採取の対象を変更することなくただ手法を適用するだけで，正解数の向上を確認した．

今後，順位づけ手法による違いや，そのエラー訂正能力，比較関係を帰納することによる細分化の効果について調べる予定である．

参考文献

[1] J. Cussens, "Part-of-Speech Tagging Using Progol" *ILP '97*, 1997.

[2] T. Nakano et al, "Inducing Shogi Heuristics Using Inductive Logic Programming" *ILP '98*, 1998.