文型パターンにおける時制・相・様相表現の汎化とその効果

金澤佑哉　　徳久雅人　　村上仁一　　池原悟
鳥取大学工学部知能情報工学科
{kanazawa,tokuhisa,murakami,ikehara}@ike.tottori-u.ac.jp

1　はじめに

等価的類推思考の原理に基づく言語の等価変換を実現するために，文型パターン辞書の構築が進められている [1]．文型パターンは，言語表現を，字面，変数，関数，記号で記述したものであり，パターンマッチングにより入力文を解析する．現在，[1] で構築されているパターン辞書では，パターンに記述された「時制・相・様相」が入力文と一致しなければパターンが適合しない．

そこで，本稿では，関数で記述された「時制・相・様相」の汎化を行い，その効果を定量的に評価することを目的とする．

2　文型パターンにおける時制・相・様相の汎化

2.1　文型パターン

文型パターンは，日英文対応の対訳コーパスから作成される．原文は，単語レベル，句レベル，節レベルの 3 レベルにパターン化され，各レベルに応じた粒度でアライメントがとれた部分は，線形要素として変数化される．逆に，変数化すると対訳の訳出が困難になる部分は変数化せず，非線形要素として字面，あるいは関数の形式で残される．

文型パターンの具体例を以下に示す．

- 日本語原文=彼は不幸にも金をなくした。
- 英語原文=He had the misfortune to lose his money.
- 日本語パターン $=N1$ は不幸にも $N2$ を $V3.kako$。
- 英語パターン $=N1$ had the misfortune to $V3$ $N1$.poss $N2$.

N は名詞，V は動詞，$.kako$ は時制を表している．英語パターンは日本語パターンを元に，英語原文を変数化して作成されている．

本稿では，[1] の文型パターンのうち文法・単語レベルの文型パターン辞書 (122,619 パターン) を「基準パターン辞書」とする．この辞書を対象に汎化を行い，新たなパターン辞書「対象パターン辞書」を作成する．

2.2　汎化の実施

2.2.1　時制関数の汎化

時制関数とはパターンに時制の情報を与える関数であり，過去関数「$.kako$」と未来関数「$.darou$」の 2 種類がある．時制関数は文型パターンの述部に記述され，そのパターンに〈過去〉または〈未来〉の時制を与えている．時制関数が文型パターンの述部に記述されていないパターンは〈現在〉の時制であることが多い．

異なる時制への適合について，その汎化の効果を得るため，過去関数と未来関数を統合し，両時制に適合する新たな関数である過去未来関数「$.kakodarou$」を定義する．そして，過去未来関数に任意記号（照合ではその要素が入力日本文にあってもなくてもよいことを示す記号，[…] で表記する）を付けることにより現在時制にも適合するようにする. 本稿ではこの関数を「自由時制関数」と呼ぶ.

自由時制関数をパターンに付加する方法は 2 つある．1 つは，時制関数を自由時制関数に置換する方法である．もう 1 つは，時制関数と様相関数が組み合わさって記述されている述部の中で，時制関数が記述されている位置を調べ，その位置を基準として時制表現の許される所 [2] に自由時制関数を挿入する方法である．こうして，時制関数を汎化したパターン辞書（自由時制パターン辞書）を作成する．

以下に具体例を示す．

方法 1

（汎化前） $N1$ が $V2.kako$。

（汎化後） $N1$ が $V2\sharp1[.kakodarou]$。

方法 2

（汎化前） $N1$ が $V2.teiru$。

（汎化後） $N1$ が $V2.teiru\sharp1[.kakodarou]$。

2.2.2　相・様相関数の汎化

相・様相関数とはパターンに相および様相の情報を与える関数であり，37 種類が定義されている．相・様相関

数も文型パターンの述部に記述され，そのパターンに相，様相の情報を与えている．

異なる相，様相への適合ついて，その汎化の効果を得るため，パターン辞書中の相・様相関数の使用頻度を調べ，頻度が高い関数について，汎化を行う．表 1 に相・様相関数と使用頻度を示す．使用頻度が 1,000 以上の関数は 10 種類あり，これらの関数に任意記号を付けて汎化する．

表 1: 相・様相関数

順位	表記	使用頻度
1	*.teiru*	9,565(17.94%)
2	*.rerurareru*	7,378(13.84%)
3	*.dadantei*	7,166(13.44%)
4	*.hitei*	6,731(12.63%)
5	*.teinei*	4,302(8.07%)
6	*.meireigotekudasai*	2,804(5.26%)
7	*.you*	2,734(5.13%)
8	*.suiteirashii*	1,779(3.34%)
9	*.sase*	1,199(2.25%)
10	*.tekuru*	1,071(2.01%)
11	*.tai*	961(1.80%)
12	*.tekureru*	954(1.79%)
13	*.joutaihenka*	891(1.67%)
14	*.teshimatta*	724(1.36%)
15	*.gimu*	574(1.08%)
その他		6,477(12.15%)
合計		55,310(100.00%)

本稿では，相・様相関数の個別の効果を調べるために，それぞれ汎化した関数ごとに新パターン辞書を作成する．また，これらの関数を同時に汎化した場合についても調べるため，使用頻度が 1,000 以上の相・様相関数に対し同時に汎化したパターン辞書も作成する．こうして，11 個のパターン辞書を作成する．

以下に複数の相・様相関数に対し同時に汎化した時の具体例を示す．

（汎化前） $N1$ は $V2.teiru.hitei$。

（汎化後） $N1$ は $V2$♯1[$.teiru$]♯2[$.hitei$]。

3 被覆率調査における評価パラメータ

本稿では，[3]，[4] で示されている評価パラメータを使用する．以下に，概略を説明する．

3.1 評価パラメータの種類

(1) 文型再現率 $R1$

$R1$ は「全入力文のうち，適合文型パターンが存在した入力文の割合」を表し，以下の式で定義される．

$$R1 = M/I \tag{1}$$

I：テスト用入力文の数

M：「自己文型パターン」以外の適合文型パターンが 1 つ以上存在した入力文の数

(2) 平均適合パターン数 N

N は「入力文に対する適合文型パターン数の平均値」を表し，以下の式で定義される．

$$N = P/I \tag{2}$$

P：適合文型パターン数

(3) 文型パターン拡大率 η

η は「評価対象の対象パターン辞書が基準パターン辞書の文型パターン数に換算して，何倍に相当するか」を表し，以下の式で定義される．

$$\eta = X/B \tag{3}$$

B：基準パターン辞書の文型パターン数

X：対象パターン辞書の文型パターン数の換算値

3.2 評価パラメータの測定方法

η を用いるには，基準パターン辞書の文型パターン数と対象パターン辞書の文型パターン数の換算値が必要となる．本稿では「文型再現率 $R1$」からみた「文型パターン拡大率 η_{R1}」と「平均適合パターン数 N」からみた「文型パターン拡大率 η_N」を用いる．文献 [4] によると，以下の特性になることがわかっている．

$$R1 = (1 - exp(-\lambda_1(p_{R1})^{\lambda_2})) * 100.0(\%) \tag{4}$$

$$N = \lambda_3 p_N \tag{5}$$

そこで，基準パターン辞書の実測値より，λ_1，λ_2，λ_3 を求め，対象パターン辞書の実測値 $R1$，N より，p_{R1} および p_N を求める．そして，(3) 式にそれぞれ代入することで，η_{R1} および η_N を求める．

4 汎化の効果の実験

4.1 実験の目的

文型パターン辞書は汎化前のパターン辞書（基準パターン辞書）と，第 2 章で作成した汎化後のパターン辞書（対象パターン辞書）を使用し，汎化の効果を第 3 章の評価パラメータを用いて定量的に評価する．

4.2 実験の方法

実験には，実際に文を入力して，パターンの被覆率を求める．テスト用の日本語入力文は 123,016 文 (=I) を使用する．入力文と文型パターン辞書を文型パターンパーサ [5] を用いて照合を行い，照合結果から，以下の項目 (i)〜(vi) を調べる．

表 2: 評価パラメータの結果

パターン辞書	M^{*1}	適合文型パターン数 P	文型再現率 $R1$	平均適合パターン数 N	$R1$ からみた η η_{R1}	N からみた η η_N	η_d^{*2}
基準パターン	59,338	1,649,455	48.24	13.41	1.00	1.00	1.00
(1) 時制関数を汎化	66,900	2,244,275	54.38	18.24	1.53	1.36	2.65
(2)[$.teiru$]	62,322	1,818,128	50.66	14.78	1.19	1.10	1.08
(3)[$.rerurareru$]	60,383	1,775,492	49.09	14.43	1.06	1.08	1.06
(4)[$.dadantei$]	59,671	1,676,581	48.51	13.63	1.02	1.02	1.06
(5)[$.hitei$]	60,074	1,679,508	48.83	13.65	1.05	1.02	1.06
(6)[$.teinei$]	59,868	1,697,710	48.67	13.80	1.03	1.03	1.04
(7)[$.meireigotekudasai$]	60,107	1,674,613	48.46	13.61	1.05	1.02	1.02
(8)[$.you$]	59,696	1,678,091	48.53	13.64	1.02	1.02	1.02
(9)[$.suiteirashii$]	60,173	1,659,082	48.91	13.49	1.07	1.01	1.01
(10)[$.sase$]	59,411	1,728,212	48.30	14.05	1.01	1.05	1.01
(11)[$.tekuru$]	59,402	1,677,447	48.29	13.64	1.01	1.02	1.01
(12) 出現頻度が 1,000 以上の相・様相関数を同時に汎化	66,361	2,304,239	53.95	18.73	1.49	1.40	1.48
(13) 時制関数と出現頻度が 1,000 以上の相・様相関数を同時に汎化	73,231	3,774,328	59.53	30.68	2.15	2.29	4.75

*1 「自己文型パターン」以外の適合文型パターンが 1 つ以上存在した入力文の数

*2 「時制・相・様相」の パターン辞書上の使用頻度からみた η

(i) アンマッチ：文法・単語レベル 123,016 文中どのパターンにもマッチしなかった文数

(ii) 自己パターンにのみマッチ：テスト文から作成された文型パターンにのみマッチした文数

(iii) 他パターンにのみマッチ：テスト文から作成された文型パターン以外の文型パターンにのみマッチした文数

(iv) 自己パターン他パターン共にマッチ：テスト文から作成された文型パターンとテスト文から作成された文型パターン以外の文型パターンが共にマッチした文数

(v) 完全一致パターン数：入力文のすべての要素が文型パターンの要素と適合する文型パターンの総数

(vi) 部分一致パターン数：入力文の一部の要素が文型パターンに定義されない要素となる文型パターンの総数

「自己文型パターン」以外の適合文型パターンが 1 つ以上存在した入力文の数 M は (iii) と (iv) を足したもので，適合文型パターン数 P は (v) と (vi) を足したものである．文型再現率 $R1$ と平均適合パターン数 N の実測値は，それぞれ (1) 式，(2) 式にテスト用入力文の数 I，M，P の実測値を代入して求める．

[4] より，基準パターン辞書の被覆率のパラメータ $R1$ および N の近似式の係数は，以下のようになる．

$$\lambda_1 = 0.0050 \quad (6)$$

$$\lambda_2 = 0.4171 \quad (7)$$

$$\lambda_3 = 13.40/122,619.0 \quad (8)$$

したがって，$R1$ および N の実測値と近似式の係数を (4) 式，(5) 式に代入し，p_{R1} および p_N を求め，p_{R1} および p_N を X とし，(3) 式より η_{R1} と η_N を求める．

4.3 実験結果

実験の結果を表 2 にまとめる（η_d については考察で述べる）．表 2 より，汎化の効果が最も高いものは辞書 (13) の時制・相・様相を同時に汎化したパターン辞書だとわかった．

5 考察

5.1 η の積の関係

表 2 より以下のことが判明した．辞書 (2) から辞書 (11) を同時に汎化したものが辞書 (12) である．辞書 (1) と辞書 (12) を同時に汎化したものが辞書 (13) である．η_{R1}，η_N はそれぞれ辞書 (2) から辞書 (11) の η の積をとると辞書 (12) の η に近い．辞書 (1) と辞書 (12) の η の積をとると辞書 (13) の η に近い．これは，汎化された部分がそれぞれのパターン辞書において独立しているためだと考えられる．

これより，それぞれの時制・相・様相関数を個別に汎化し，それぞれの汎化の効果の積をとれば，時制・相・様相関数を同時に汎化したパターン辞書の汎化の効果が予測できると考えられる．

5.2 パターン辞書を汎化しない推定法

η_{R1} と η_N のように近似的に対象辞書を構築せず，汎化の効果を予想する推定法を検討する．そこで，本稿では，「時制・相・様相」のパターン辞書上の使用頻度に関

して求めた η_d を提案する．

η_d を求めるには，1 パターンに対する汎化した箇所を求める．例えば，1 パターンに対し 1 箇所汎化をした時，そのパターンは 2 倍に増えたことになる．また，2 箇所汎化をした時は，そのパターンは 4 倍に増えたことになる．

以下に 2 箇所汎化した時の具体例を示す．

（汎化前） $N1$ が $V2.teiru.hitei$。

（汎化後） $N1$ が $V2\sharp1[.teiru]\sharp2[.hitei]$。

（汎化後のパターンの展開 1） $N1$ が $V2$。

（汎化後のパターンの展開 2） $N1$ が $V2.teiru$。

（汎化後のパターンの展開 3） $N1$ が $V2.hitei$。

（汎化後のパターンの展開 4） $N1$ が $V2.teiru.hitei$。

このように，汎化したことによって増えたパターンの総数を求め，その総数を基準パターンのパターン数で割ったものを η_d とする．それぞれのパターン辞書の η_d の値を表 2 にまとめる．

表 2 より，辞書 (2) から辞書 (12) の η_{R1}，η_N の値は η_d の値と近似している．しかし，辞書 (1) と辞書 (13) の η_{R1}，η_N は η_d と値がかなり異なっている．これは，可能な限り時制を汎化したので，入力文と照合しないパターンが多数存在したためだと考えられる．

これより η_d は相・様相の汎化では汎化の効果の予測ができる．しかし，時制の汎化では汎化の効果の予測はできないと考えられる．

5.3 規模の異なるパターン辞書の汎化による推定法

大規模な被覆率調査を行うと大量の作業コストがかかる．そこで，規模の異なる被覆率調査行い，あらかじめ汎化の効果を予測する推定法を検討する．大規模な被覆率調査で求めた η と，規模の異なる被覆率調査で求めた η を比較し，規模の異なる被覆率調査でも，大規模な被覆率調査を行ったときと同等の汎化の効果を予測できるか確認する．

本稿では，規模の異なる自由時制パターン辞書を対象パターン辞書とし，規模の異なるパターン辞書ごとに η を算出し，自由時制パターン辞書の全パターン数の η と比較を行う．それぞれの η の値を表 3 にまとめる．

その結果，80,000 パターンの η が全パターン数の η に最も近い値となった．ただし，小数点第 1 位の誤差を許せば，40,000 パターンでも汎化の効果が予測できると考えられる．100 から 20000 パターンでは，データが大幅に変動するので，値に信頼性が持てない．η_d はパターン数が変動しても値が安定しているが，5.2 で述べたように，時制の汎化に関しては汎化の効果は予測ができない．

表 3: 規模の異なる自由時制パターン辞書の η

小規模パターン数	規模の異なるパターン数の		
	η_{R1}	η_N	η_d
100	0.01	0.15	2.72
500	0.13	0.15	2.82
1,000	4.07	1.12	2.64
5,000	1.64	1.30	2.63
10,000	1.83	1.40	2.70
20,000	2.05	1.34	2.67
40,000	1.72	1.47	2.69
80,000	1.69	1.39	2.67
全パターン数	全パターン数の		
	η_{R1}	η_N	η_d
122,619	1.53	1.36	2.65

これより，規模の異なる被覆率調査で求めた実測値 η も汎化の効果の予測ができる可能性があることがわかった．今後様々に考えられる汎化の方法の中で，効果の見込まれる汎化の方法の選別に有効である．

6 おわりに

本研究では「要素の位置変更可能の指定」,「時制・相・様相を表す関数の汎化」をする場合としない場合についての被覆率調査を行い，η を用いて定量的に評価した．実験結果より,「時制・相・様相を表す関数の汎化」をすることは，パターンの汎用性を向上するために有効だとわかった．また，汎化の効果があらかじめ予想できることを確認した．

今後は，各 η の異なりが生じる原因を定性的に考察し，単語レベルだけではなく，句レベル，節レベルの文型パターンの汎化を行い，汎化の効果を調査する予定である．

参考文献

[1] 池原悟, 阿部さつき, 徳久雅人, 村上仁一:非線形な表現構造に着目した重文と複文の日英文型パターン化, 自然言語処理,11(3),pp.69-95,2004.

[2] 南不二男:現代日本語文法の輪郭, 大修館書店,1993.

[3] 池原悟, 徳久雅人, 竹内 (村本) 奈央, 村上仁一:日本語重文・複文を対象とした文法レベル文型パターンの被覆率特性, 自然言語処理,11(4),pp.147-178,2004.

[4] 遠藤久美子, 徳久雅人, 村上仁一, 池原悟:文型パターンにおける任意要素の記述方法とその効果, 言語処理学会第 11 回年次大会,2005(発表予定).

[5] 徳久雅人, 池原悟, 村上仁一:文型パターンパーサの試作, 言語処理学会第 10 回年次大会発表論文集,pp.608-611,2004.