記号と関数による文型パターン汎化効果の総合評価

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

パターンマッチングにより入力文を解析する方式に対する課題は，「できるだけ多くの文がパターンに適合すること」，すなわち，「パターンの網羅性を確保すること」である．そのため，パターンを大量に作成することと，パターンの汎用性を高めることが要求される．パターンに基づく知識ベースを開発する上で，パターン数を増やすコストや，汎用性の高いパターンの記述を行うコストは，実質的に重要なファクタであるため，知識ベースの開発事例における実測値に対する関心は高い．

本稿では，重文・複文を対象とした文型パターン辞書 [1] の開発において，パターンの汎用性を高めるためのパターンの記述子である「記号」と「関数」を紹介し，その「汎用性の評価」，および，「汎化作業コスト」について報告する．

パターンの汎用性の評価は，単純にいえばパターンの網羅性（ある文集合におけるパターンのマッチした文の割合）で評価できるが，本稿では，パターン数の増加に喩えて判断できる評価パラメータを用いる [2]．汎化作業コストは，パターン辞書の組織的な開発過程におけるアナリストやプログラマの作業工数をベースに算出する．

2　文型パターンにおける記号・関数

文型パターンは，重文・複文を対象とした日英文対応の対訳コーパスから作成される．日本語文は，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，変数化される．ただし，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，字面，あるいは関数の形式で残される．文型パターンの例を以下に示す．

- 日本語原文= こうしてふたりの間にロマンスが芽生えた。
- 英語原文　= Thus a romantic relationship was formed between the two people.
- 日本語 P　= */y* こう!して *N*1 の間に *N*2 が *V*3.*kako*。
- 英語 P　= Thus *N*2 *V*3^*past*^*passive* between *N*1.

N は名詞，*V* は動詞，*/y* は離散記号，*.kako* は時制を表している．

本稿で取り扱う記号，関数について説明する．記号は，文型パターン全体の非線型な構造を保ちながら，構造上の線型性を記述するために用いられており，表 1 の 6 種類を取り扱う．

表 1　記号一覧

記述子名	記述形式	意味
離散記号	*/y*：連用節，*/t*：連体節，*/c*：格要素，*/f*：副詞，*/k*：形容（動）詞連体形・連体詞	文型パターンにない要素で，挿入可能なものを指定（5 種類）
文節境界記号	!	文節の境界を受理
要素選択記号	(…|…)	いずれかの要素列を受理
任意要素記号	[…]	文型選択上，任意の要素
順序任意記号	{…,…}	順序入れ替え可能な範囲
位置変更可能記号	$*n*^{…}, $*n*	指定位置に入れ替え可能

関数は，表記上の揺らぎを吸収し，適用範囲の広い文型パターンを記述するために用いられており，表 2 の 2 種類を取り扱う．

表 2　関数一覧

記述子名	記述形式	意味
日本語様相関数	. 関数名（*.kako*：過去，*.tearu*：てある，*.you*：意志，…）	時制，相，態，様相などの助動詞相当表現を指定（39 種類）
同値型グループ関数	.# 関数名（*.#da*：体言述部型，*.#darou*：推量型）	同種の意味の関数をグループ化して指定（2 種類）

3　評価パラメータ

3.1　網羅性の評価

網羅性を評価するパラメータとして，「文型再現率 *R*1」，「適合文型数 *N*」を使用する [2].

3.1.1 文型再現率 $R1$

$R1$ は「全入力文のうち，適合する文型パターン（以下，適合パターンと呼ぶ）の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：全入力文の数
M：適合パターンが 1 つ以上存在した入力文の数

3.1.2 適合文型数 N

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

$$N = \sum_{i=1}^{M} N_i/M \tag{2}$$

N_i：i 番目の入力文の適合パターンの数

3.2 パターンの増加数への換算

文型再現率 $R1$ や適合文型数 N は，文型パターン辞書の規模（パターン数）に対して単調増加する．パターン数 p の文型パターン辞書における $R1$ の曲線は次式で表される [3]．なお，[1] の文型パターン辞書において，$\lambda_1 = 0.00407944$，$\lambda_2 = 0.47791782$ である．

$$R1 = (1-exp(-\lambda_1(p)^{\lambda_2})) \times 100.00(\%) \tag{3}$$

[1] の文型パターン辞書の記述子に変更を加えると，入力文と文型パターンとの適合の仕方が変化して，文型再現率 $R1$ や適合文型数 N が変化する．式 (3) から p を表す逆関数を求め，その変化後の $R1$ を代入すると，記述子に変更を加えられた文型パターン辞書の規模の換算値を求めることができる．

3.3 文型パターン拡大率 η

文型パターン拡大率 η は「評価対象のパターン辞書（以下，対象パターン辞書と呼ぶ）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．

$$\eta = X/B \tag{4}$$

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

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}」を使用する．このとき，換算値 X は式 (3) から求める．

4 汎用性の評価

4.1 汎用性の計測方法

[1] の文型パターン辞書（217,703 パターン）から，文法・単語レベルの文型パターン集（119,229 パターン）を抽出し，それを「基準パターン辞書 ❶ 」として，記号・関数の汎用性を評価する．

基準パターン辞書には，既に記号・関数が導入されている．記号・関数の各種類ごとに汎用性向上への寄与を評価するために，基準パターン辞書から，記号・関数を削除したパターン辞書を作成して，削除パターン辞書の文型パターン拡大率 η を計測する．汎化機能のある記述子を削除するため，η は 1 より小さい場合に汎化の寄与があるといえる．

計測には，実際に入力文と文型パターンを照合して，第 3 章で示した評価パラメータを求める．計測用の入力文は，対訳標本からランダムに選んだ 10,000 文 (= I) とする．文型パターンを作成するために用いた文であるが，入力文から作られた文型パターンは，照合に用いないので，クロスバリデーション型の実験である．入力文に適合する文型パターンを辞書から検索するには，検索ツール（SPM）を用いる [4]．

4.2 汎用性の評価結果

出現回数（基準パターン辞書に出現した記号・関数の回数）と評価パラメータの計測結果を表 3 にまとめる．表 3 から，以下のことが分かる．

(1) 記号のうち，汎化効果が最も高いものは，辞書①の離散記号である．辞書①の離散記号の有無による違いを比較すると，以下のことが分かる．
- $R1$ が 37.99% 向上（辞書 ❶ の $R1$– 辞書①の $R1$）．
- N が 22.34 向上（辞書 ❶ の N– 辞書①の N）．
- 文型パターン数が 14.29 倍増加（辞書 ❶ の η_{R1}÷ 辞書①の η_{R1}）．
- 離散記号の出現回数は記号の中で最も多く，1 パターンに平均 6.38 回出現（辞書①の出現回数 ÷ 辞書 ❶ のパターン数）．

(2) 関数のうち，汎化効果が最も高いものは，辞書⑧の日本語様相関数である．辞書⑧の日本語様相関数の有無による違いを比較すると，以下のことが分かる．
- $R1$ が 26.34% 向上（辞書 ❶ の $R1$– 辞書⑧の $R1$）．
- N が 8.55 向上（辞書 ❶ の N– 辞書⑧の N）．
- 文型パターン数が 5.26 倍増加（辞書 ❶ の η_{R1}÷ 辞

表 3　出現回数と評価パラメータの計測結果

パターン辞書＼評価項目	出現回数	$R1$(%)（適合文数 / 入力文数）	N（適合パターン数 / 適合文数）	換算値 X（パターン数）	η_{R1}
❶ 基準パターン辞書	-	60.18 (6,018/10,000)	36.71 (220,925/6,018)	119,229	1.00
① 離散記号を削除した辞書	760,145	22.19 (2,219/10,000)	14.37 (31,882/2,219)	8,346	0.07
② 文節境界記号を削除した辞書	70,621	60.23 (6,023/10,000)	36.71 (221,106/6,023)	119,229	1.00
③ 要素選択記号を削除した辞書	158,764	59.21 (5,921/10,000)	32.07 (189,862/5,921)	113,268	0.95
④ 任意要素記号を削除した辞書	53,475	54.38 (5,438/10,000)	35.07 (190,691/5,438)	85,845	0.72
⑤ 順序任意記号を削除した辞書	13,260	60.10 (6,010/10,000)	36.53 (219,523/6,010)	118,037	0.99
⑥ 位置変更可能記号を削除した辞書	35,416	59.90 (5,990/10,000)	34.42 (206,180/5,990)	116,844	0.98
⑦ 全ての記号を削除した辞書	1,204,785	18.48 (1,848/10,000)	11.71 (21,642/1,848)	4,769	0.04
⑧ 日本語様相関数を削除した辞書	120,633	33.84 (3,384/10,000)	28.16 (95,282/3,384)	22,654	0.19
⑨ 同値型グループ関数を削除した辞書	10,380	57.87 (5,787/10,000)	36.35 (210,349/5,787)	104,922	0.88
⑩ 全ての関数を削除した辞書	131,013	30.55 (3,055/10,000)	25.89 (79,088/3,055)	16,692	0.14
⑪ 全ての記号・関数を削除した辞書	1,335,798	8.90 (890/10,000)	8.21 (7,303/890)	1,192	0.01

~~~~~は ❶ より値が低いとき，＝＝＝＝＝は ❶ より値が高いとき付与

書⑧の η_{R1}）．

- 日本語様相関数の出現回数は関数の中で最も多く，1 パターンに平均 1.10 回出現（辞書⑧の出現回数 ÷ 辞書 ❶ のパターン数）．

(3) 辞書⑪の全ての記号・関数の有無による違いを比較すると，以下のことが分かる．

- $R1$ が 51.28% 向上（辞書 ❶ の $R1$– 辞書⑪の $R1$）．
- N が 28.50 向上（辞書 ❶ の N– 辞書⑪の N）．
- 文型パターン数が 100.00 倍増加（辞書 ❶ の η_{R1}÷ 辞書⑩の η_{R1}）．
- 記号・関数は，1 パターンに平均 11.20 回出現（辞書⑧の出現回数 ÷ 辞書 ❶ のパターン数）．

これらから，記号・関数が，パターンの汎用性を高めていることを確認できた．

5　汎化作業コスト

5.1　汎化作業の関係

記号・関数を文型パターンに導入する際，日英対訳文対における統語的・意味的な関係をアナリストが分析し，導入可能であるかどうかを検査している．その前後には実務的な作業が関わっている．基礎的な作業として，(a) ツール作成，(b) データ抽出，(c) 実作業（分析作業），(d) 作業結果の反映という段階がある．(c) はツール上で作業する場合もあれば，紙の上での作業と入力作業が混在する場合もある．(d) は，中枢のデータベースに作業結果の形式を整えつつ登録する部分がある．

文型パターンの記述子ごとに参照する言語的情報の要点がある．具体的に基礎となる言語的情報を表 4 の大項目と小項目に示し，その（a）〜（d）の作業コストを人日の単位で示す．

また，記述子ごとの作業では，この言語的情報を，部分的かつ混合して使用するため，作業工数の算出は単純ではないが，表 5 の「小計」の列の通りとなった（作業コストについては 5.2 節で述べる）．

例えば，表 5 によると，離散記号の導入は，「日本語形態素の品詞等の情報（2-1）」，および，「日本語節間キーワード（3）」を基礎情報として使用し，ある程度自動的に文型パターンに導入していることが（d）で読み取れる．

5.2　記号・関数に対する作業コスト

記号・関数に対する「作業コスト」は，表 5 の「小計」と「使用した基礎情報（表 4 から算出）」の和で算出する．例えば，離散記号の作業コストは，「小計」+「使用した基礎情報」= 40 + (105 + 80) = 225 人日となる．表 5 から，以下のことが分かる．

(1) 汎化作業コストが最も低いのは，文節境界記号の 195 人日である．
(2) 汎化作業コストが最も高いのは，関数の 1,125 人日である．

これらから，作業コスト間で，最大 6 倍程度の差があることが分かった．

6　総合評価

第 4 章で求めた「汎用性の評価」と第 5 章で求めた「汎化作業コスト」から，1 人日あたりの作成パターン数に換算した総合評価を行なう．算出方法は，表 3 から基準
~~~~~

表 4　基礎となる言語的情報の作業工数

大項目	小項目	(a)	(b)	(c)	(d)	小計
1 日英対応付け作業	変数化，補完主語，英語時制等の対応付け	40	20	620	20	700
2 日本語形態素修正	1 品詞，単語・文節境界等の修正	30	10	60	5	105
	2 標準表記の修正		5	20	5	30
3 日本語節間キーワード付与				80		80

（単位 = 人日）

表 5　記述子ごとの導入作業の工数と作業コスト

記述子類	記述子名	使用した基礎情報	(a)	(b)	(c)	(d)	小計	作業コスト
記号	離散記号	2-1，3				40	40	225
	文節境界記号	2-1，3				10	10	195
	要素選択記号	2	5	10	60	20	95	230
	任意要素記号	1					0	920
	順序任意記号	2-1	30	15	140	10	195	300
	位置変更可能記号	2-1	30	15	180	10	235	340
関数		1，2-1				100	100	1,125

（単位 = 人日）

パターン辞書のパターン数と換算値 X の差を求め，表 5 の作業コストで割ることで算出する．ただし，単語レベルの作業コストに換算するため，作業工数に 0.55（単語レベルの割合）を掛けた値を使用する．

例えば，離散記号の場合は，「文型パターン作成数」÷「作業コスト」= (119,229 − 8,346) ÷ (225 × 0.55) = 896 となる．表 6 に，1 人日あたりの作成パターン数が多い順の総合評価結果を示す．

表 6　総合評価結果

順位	記述子名（類）	1 人日あたりの文型パターン作成数	（文型パターン作成数 / 作業コスト）
1	離散記号	896	(110,883 / 123.75)
2	関数	166	(102,537 / 618.75)
3	任意要素記号	66	(33,384 / 506.00)
4	要素選択記号	47	(5,961 / 126.50)
5	位置変更可能記号	13	(2,385 / 187.00)
6	順序任意記号	7	(1,192 / 165.00)

表 6 から，以下のことが分かる．

(1) 総合評価が最も良いのは，離散記号で，少量の作業コストで，文型パターンが大量に作成できる．
(2) 関数は，作業コストが大量にかかるが，文型パターンも大量に作成できる．

また，表 5 から，順序任意記号と位置変更可能記号は，記述子導入に関わる共通部分を含まない作業コスト（表 5 の「小計」）が大量にかかることが分かる（195 人日，235 人日）．しかし，表 6 から，1 人日あたりの作成パターン数にはそれほど効果がなかったことが分かった．

7　おわりに

本稿では，パターンの汎用性を高めるためのパターンの記述子である「記号」と「関数」について，その汎用性の評価，および，汎化作業コストを求めて，総合評価を行った．

その結果，総合評価が最も良いのは，離散記号で，1 人日あたり 896 パターン相当作成出来ることが分かった．次点の関数は，種類数が多く作業コストがかかったが，汎化の効果が高く，導入は妥当であると言える．

謝辞

本研究は，独立行政法人科学技術振興機構（JST）・戦略的創造研究推進事業（CREST）の研究領域「高度メディア社会の生活情報技術」の研究課題「セマンティックタイポロジーによる言語の等価変換と生成技術」の支援によるものである．

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] 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.
[3] 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, pp.368-371, 2005.
[4] 徳久雅人, 村上仁一, 池原悟:重文・複文文型パターン辞書からの構造照合型パターン検索, 情報処理学会研究報告, 自然言語処理, 2006-NL-176, pp.9-16, 2006.