052009 記号と関数による文型パターン汎化効果の総合評価

計算機工学講座 C　金澤 佑哉

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．

本研究では，文型パターン辞書 [1] 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，単にパターンの網羅性（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，本研究では，パターン数の増加に比例した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．なお，*N* は名詞，*V* は動詞，*/y* は離散記号，*.kako* は時制関数を表している．

- 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.

本研究で取り扱う記号，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つつ，構造上の線型性を記述するために用いる．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る．表 1 に，取り扱う記号（6 種類①〜⑥）と関数（2 種類⑧，⑨）の「記述形式」と「意味」を示す（評価パラメータの計測結果は 4.2 節で述べる）．

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」を使用する [2]．*R*1 は以下の式で定義される．

$$R1 = \frac{\text{適合するパターンが 1 つ以上ある入力文の数}}{\text{全入力文の数}} \quad (1)$$

3.2　パターンの増加数への換算

*R*1 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 *R*1 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．

$$R1 = (1 - exp(-\lambda_1 (p)^{\lambda_2})) \times 100.00(\%) \quad (2)$$

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，*R*1 が変化する．式 (2) から p を表す逆関数を求め，その変化後の *R*1 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．

$$\eta = \frac{\text{基準パターン辞書の文型パターン数}}{\text{対象パターン辞書の文型パターン数の換算値 } (=X)} \quad (3)$$

本研究では，「文型再現率 *R*1」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 (2) 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 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，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

評価パラメータの計測結果を表 1 の「*R*1」,「換算値 *X*」,「η_{R1}」にまとめる．表 1 から，以下のことが分かる．

表 1 記号・関数の説明と評価パラメータの計測結果

記述子類	項目 / 記述子名（辞書）	記述形式	意味	$R1$(%)	換算値 X	η_{R1}
	❶ 基準パターン辞書	-	-	60.18	119,229	1.00
記号	① 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）	22.19	8,346	0.07
記号	② 文節境界記号	!	文節の境界を受理	60.23	119,229	1.00
記号	③ 要素選択記号	(…|…)	いずれかの要素列を受理	59.21	113,268	0.95
記号	④ 任意要素記号	[…]	文型選択上，任意の要素	54.38	85,845	0.72
記号	⑤ 順序任意記号	{…,…}	順序入れ替え可能な範囲	60.10	118,037	0.99
記号	⑥ 位置変更可能記号	$*n*^{…}, $*n*	指定位置に入れ替え可能	59.90	116,844	0.98
記号	⑦ 全ての記号	-	-	18.48	4,769	0.04
関数	⑧ 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）	33.84	22,654	0.19
関数	⑨ 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）	57.87	104,922	0.88
関数	⑩ 全ての関数	-	-	30.55	16,692	0.14
	⑪ 全ての記号・関数	-	-	8.90	1,192	0.01

波線は ❶ より値が低いとき，二重線は ❶ より値が高いとき付与

(1) 記号では，離散記号が最も効果があり，文型パターン数が 14.29 倍増加した（辞書 ❶ の η_{R1}÷ 辞書①の η_{R1}）.
(2) 関数では，日本語様相関数が最も効果があり，文型パターン数が 5.26 倍増加した（辞書 ❶ の η_{R1}÷ 辞書⑧の η_{R1}）.

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

記号・関数を，単語・句・節レベルの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．知識ベースの開発事例における，記号・関数に対する「作業コスト」を表 2 に示す.

表 2 記号・関数に対する作業コスト

（単位＝人日）

	パターン辞書	作業コスト
記号	離散記号	225
	文節境界記号	195
	要素選択記号	230
	任意要素記号	920
	順序任意記号	300
	位置変更可能記号	320
関数		1,125

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 1 から基準パターン辞書のパターン数と換算値 X の差を求め，表 2 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．表 3 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 3 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 3 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，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.

7 おわりに

本研究では，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．これらから，パターンに基づく知識ベースの開発における，コストの判断に有効な評価パラメータを見出すことが出来た.

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.
[5] 金澤佑哉, 徳久雅人, 村上仁一, 池原悟: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, 言語処理学会第 11 回年次大会, pp.29-32, 2005.
[6] 金澤佑哉, 徳久雅人, 村上仁一, 池原悟:記号と関数による文型パターン汎化効果の総合評価, 言語処理学会第 13 回年次大会, 2007（発表予定）.