012012 日本語文型パターンの縮退方法

計算機工学講座 池原研究室　片山慶一郎

1　はじめに

現在，24 万件の文型パターン辞書がある [1]．この文型パターン辞書は，15 万件の日英対訳コーパスを元に作成されている．規模が大きく取り扱いが困難である為，パターン数を削減する方法が必要とされているが，人手によりパターン数を削減する事は困難である．

そこで，本研究では日本語文型パターン間の包含関係に着目して，その数を半自動的に削減する方法を検討する．

2　パターンの削減方法

2.1　パターン間の包含関係

パターン $P1$ に適合する入力文の全てがパターン $P0$ に適合する時，パターン $P1$ はパターン $P0$ に包含されると定義し，また，パターン $P0$ を上位のパターン，パターン $P1$ を下位のパターンと呼ぶ．両者の関係を $P0 \supseteq P1$ と表記する．

以下に包含関係にあるパターンの例を示す．

$P0$: $N1$ は [$N2$ を]$VP3$。
$P1$: 私は $V1$。

本研究では，パターン辞書からパターン $P1$ を削除することで，パターン数の削減を試みる．

2.2　パターン要素間の包含関係

パターン間の包含関係を考える場合，パターンを構成する要素（変数，関数，記号，字面）の包含関係を定義する必要がある．そこで，各要素の定義 [2] に基づいて包含関係を定義する．表 1 に要素間の関係の一部を示す．

表 1　パターン要素間の関係の一部

VP	$\supseteq$	$V, VP, \ldots$
N	$\supseteq$	$N, NUM, \ldots$

VP: 動詞句変数, V: 動詞変数, N: 名詞変数, NUM: 数詞変数

2.3　パターンパーサ

文型パターンパーサ (以下パーサ)[3] は，日本語文型パターンと日本語文との照合を行うプログラムである．照合方式は，ATN(Augmented Transition Network)[4] をベースとしている．

そこで，本研究においてもこのパーサを用い，パターン間の包含関係の判定を行う．

3　包含関係による削減の実験

3.1　包含関係の判定方法の実装

パーサは日本語文とパターンでの照合を想定して作成されている．その為，パターンをパーサの入力文仕様に合わせる為に以下の作業を行う．

1. 要素選択記号，任意記号など日本語文に存在し得ない記号を展開．
2. 変数等を含む文型パターンの形態素解析結果を作成．
3. 入力文としての変数・関数をパターン側の変数が表 1 で示される包含関係を用いて受理出来るように，変数・関数定義オートマトンの修正．

3.2　実験対象

文型パターン辞書に収録されている文法・単語レベルの 122,619 パターンを実験の対象パターンとする．

3.3　パーサの出力結果

パーサを用いてパターン間の包含関係を調査した．以下に出力例を示す．この例では，$P2 \supseteq P3$ となる．

$P2$: /ytk $N1$ は /tcfk $N2$ の /f ある！ $N3.da$。
$P3$: /ytk $N1$ は /tcfk 力の /f ある！ $N2.da$。

パターン $P3$ では，名詞「力」が変数に置き換えられていない．その為，パターン $P2$ と $P3$ で表記が異なっている．包含関係にあるパターンの多くは同様の理由であった．

3.4　包含関係を用いた削減結果

調査した包含関係を用い，下位パターンを削除した．この結果，パターン数は 112,767（削減率は 8.0%）となった．

3.5　出現頻度データとの関係

包含関係により削減したパターン数と，出現頻度との関係を調査した．出現頻度は，本研究で用いたパターンが適合した原文の数で，パターンと全原文との間の照合実験により求めた．調査結果を表 2 に示す．

表 2　出現頻度データと包含関係による削減結果の関係

出現頻度	パターン数	削減パターン数
1000 文以上	275 (0.2%)	195 (70.9%)
100 文以上 1000 文未満	1,013 (0.8%)	502 (49.6%)
100 文未満 10 文以上	3,194 (2.6%)	983 (30.8%)
10 文未満 2 文以上	13,140 (10.7%)	3,255 (24.8%)
小計	**17,622 (14.4%)**	**4,935 (28.0%)**
1 文以下	104,997 (85.6%)	4,917 (4.7%)
合計	**122,619 (100.0%)**	**9,852 (8.0%)**

表 2 より，以下のことが示される．(1) 出現頻度が大きいパターンは削減率が大きい．(2) 出現頻度が小さいパターンは削減率が小さい．

4　考察

4.1　パーサの判定結果の考察

パーサは，以下に示すパターン $P4$ はパターン $P2$ に一部包含されると判定した．

$P4$: /y </tk $N1$ は> /tcfk 張合の /f ある！ $N2.da$。

パターン $P4$ は，以下に示す 2 パターンに展開される．

$P4$-0: /y /tcfk 張合の /f ある！ $N2.da$。
$P4$-1: /y /tk $N1$ は /tcfk 張合の /f ある！ $N2.da$。

展開後のパターン間の包含関係は，$P2 \supseteq P4$-1, $P2 \not\supseteq P4$-0 となる．したがって，$P2 \not\supseteq P4$ と判定される．しかし，パターン $P2$ を次のパターン $P2'$ に示すように変更することで，$P2' \supseteq P4$ となる．

$P2'$: /y </tk $N1$ は> /tcfk $N2$ の /f ある！ $N3.da$。

従って，一部包含関係があると判定されたパターンに着目し，両者を含む新たなパターンを作成することで，削減率の向上を図れる可能性がある．

4.2　出現頻度データとの関係の考察

出現頻度が低いパターンでは削減率が小さかった．これは，出現頻度が低いパターンは特有の表現を持つものが多い為と考えられる．

5　おわりに

本研究では，パターン間の包含関係に着目して大規模文型パターンの削減を試みた．その結果，全体の 8% のパターンを削除出来た．また，出現頻度が高いパターンは削減率が大きいが，出現頻度が低いパターンは削減率が小さかった．

今後，包含関係には無いが，類似度の高いパターンを調査し，両者を含む新たなパターンを作成する手法が考えられる．また，削減が出来なかったパターンについては，縮退自体の必要性の有無も検討していく．

参考文献

[1] 池原ほか: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, 言語処理学会論文誌 Vol.11, No.3, pp.69-95, 2004.
[2] 池原ほか:機械翻訳のための日英文型パターン記述言語の設計, 電子情報通信学会技術研究報告,TL2002-48, pp.1-6, 2003.
[3] 徳久ほか: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.
[4] James Allen:Natural Language Understanding(2nd Edition), The Benjamin/Cummings Publishing Company, Inc., pp.101-106, 1994.