意味類型化のための名詞述語文のパターン化

藤原 竜樹 † 徳久 雅人 † 村上 仁一 † 村田 真樹 †
鳥取大学大学院工学研究科 情報エレクトロニクス専攻 †

1 はじめに

日本語の単文は，用言述語文と名詞述語文に大別される．用言述語文は，形容詞または形容動詞により属性を述べる文，および，動詞により事態を述べる文である．用言述語文に対応する意味解析用の知識ベースは既に存在する [1]．一方，名詞述語文は，主語の帰属する範疇を述べる文（範疇叙述型），内包的概念を表す主語の外延を述べる文（外延叙述型），および，主語の数量などの属性を述べる文（属性叙述型）に分類される [2]．しかし，対応する意味解析用の知識ベースは存在しない．

そこで，本稿では，名詞述語文の意味を解析するための知識ベースとしてパターンを作成することを目的とする．ここで意味解析とは，入力文の型を判定し，範疇や属性などの情報を抽出することである．

以上に向けて，第２章ではパターンの作成方法を示す．第３章では入力文の意味を解析するためのパターン集の運用方法を示す．第４章ではパターンによる解析性能を実験により評価する．第５章では考察を行う．最後に第６章でまとめを述べる．

2 パターンの作成

2.1 設計

範疇叙述型および外延叙述型は，上位語，下位語，および，両者がつりあうための追加情報で構成される．属性叙述型は，属性，属性値，および，それらを持つ実体で構成される [2]．これらの情報を抽出するため，これらに該当する部分を変数化し，汎化することでパターンを作成する．変数のマッチできる意味的な範囲を制約条件（意味制約）として付記する．ここで，意味制約に用いる意味コードは [1] を用いる．例えば，「パソコン」は 971 である．

2.2 作成の結果

文献 [2] に示された例文 37 文を基に，パターンを作成したところ，8 パターンおよび 37 通りの意味制約を得た．以下に原文と作成例を示す．MT は，体言の修飾語句を表す変数，MD はモダリティ表現を表す変数である．その他の記述子は文献 [3] を参照されたい．

例 1.　原文：パソコンはただの道具だ。
パターン：
$/[MT1]NP2(は \mid が)[、]/yf[MT3]NP4[MD5]$
型名：範疇叙述型
意味制約：$NP2(971);NP4(1035,893)$
３つ組：下位:$NP2$, 上位:$NP4$, 追加情報:$MT3$

例 2.　原文：使った道具はドライバーだ。
パターン：
$/[MT1]NP2(は \mid が)[、]/yf[MT3]NP4[MD5]$
型名：外延叙述型
意味制約：$NP2(292,942);NP4(1035,893)$
３つ組：上位:$NP2$, 下位:$NP4$, 追加情報:$MT1$

例 3.　原文：このカレーは３００円だ。
パターン：
$/[MT1]NP2(は \mid が)[、]/yf[MT3]NP4[MD5]$
型名：属性叙述型
意味制約：$NP2(848,852);NP4(2595,2590,1190)$
３つ組：実体:$NP2$, 属性:$NP4$, 属性値:$NP4$

3 パターン集の運用

3.1 運用手順

まず，入力文を全てのパターンと照合すると，型名および３つ組の情報が複数得られる．次に，3.2 節の方法で適合結果の選択を行う．最後に，選択した型名および３つ組情報を，入力文の意味解析結果として出力する．

3.2 最適な適合結果の選択

最適な適合結果の選択は，以下の手順で行う．

手順１：変数 MD がマッチした適合結果を優先して選択
手順２：追加情報および属性値のある適合結果を優先して選択
手順３：変数がより具体的な適合結果を優先して選択
手順４：主語，述語の上下関係をもとに優先して選択
手順５：原文との意味的な近さをもとに選択

ここで，手順１，２，３は，文法レベルでパターンが適切にマッチしたかを調べ，手順４，５は，意味レベルでパターンが適切にマッチしたかを調べるものである．

Patterns of noun predicate sentences to analyze meaning type
†Department of Information and Electronics, Graduate School of Engineering, Tottori University

手順１では，名詞述語文の判定詞を確認することで，述語性の確保を行う．手順２は，それぞれの型の文の構造を確認する．範疇叙述型は述語の前，外延叙述型は主語の前に追加情報があるかを確認を行う．属性叙述型では，属性値が抽出されているか確認する．手順３では，変数がより具体的な適合結果を選択することにより，文法的な具体性を確認する．手順４は，範疇叙述型および外延叙述型の意味的な文の構造を確認する．主語および述語に付与されている意味コードにより，範疇叙述型では主語が下位で述語が上位の関係，外延叙述型では主語が上位で述語が下位の関係が，それぞれ成り立っている適合結果を選択する．手順５は 3.3 節で述べる．

3.3　原文との意味的な近さ

入力文とパターンの原文が意味的に近いことを条件に，最適な適合結果を選ぶ．上位語や属性表現の意味の近さを重視しており，式 (1) で最適な適合結果 $\hat{m}$ を求める．

$$\hat{m} = \underset{\substack{m\in M \\ (C_{主},C_{述})\leftarrow m \\ c_1\in C_{i\ 主},c_2\in C_{主} \\ c_3\in C_{i\ 述},c_4\in C_{述}}}{\arg\min} \{w_1 d(c_1,c_2) + w_2 d(c_3,c_4)\} \quad (1)$$

ここで，M は適合結果の集合，m は適合結果（原文，パターン，型名，意味制約，３つ組，および，パターンマッチで代入される変数値で構成．なお，M は原文の数および代入のバリエーションにより増大），$(C_{主}, C_{述})$ は m からとり出される意味制約であり，$C_{述}$ はなかでも述語に対する集合（例１では，$\{1035, 893\}$），$C_{i\ 主}$ は入力文 i の主語の意味コードの集合（一語につき複数のコードが存在），$C_{主}$，$C_{i\ 述}$ も同様の集合，(w_1, w_2) は，m の型が外延叙述型ならば $(1.0, 0.1)$ とし，その他ならば $(0.1, 1.0)$ とする．d はシソーラス [1] において，２つの意味コードの距離を表す関数である．

シソーラスは，名詞の概念の上下関係を木構造で表す．ルートノードは最上位であり，抽象度が最も高い．下位ノードであるほど具体的である．ノードは意味コードが付与されている．アークは上下関係のあることを表す．関数 d は，２つの意味コードに対応するノード間のアーク数を返すことで，意味の距離を表す．

4　実験

クローズドテスト（入力 37 文，正解データ 37 件）およびこのデータを用いた leave-one-out cross-validation(LOOCV) を行った．再現率 R，適合率 P，および，F 値 を用いて，３つ組の抽出性能を評価した（表 1）．ここで，$R = \langle$ 一致数 $\rangle / \langle$ 入力数 $\rangle$，$P = \langle$ 一致数 $\rangle / \langle$ 出力数 $\rangle$，F 値 $= 2PR/(P+R)$ である．

表 1　意味解析の性能

テスト名	入力数	出力数	一致数	R	P	F 値
Closed	37	54	37	1.00	0.69	0.81
LOOCV	37	52	34	0.92	0.65	0.76

5　考察

手順を１から４まで処理を行った場合，または処理を行わない場合の抽出性能の評価を示す（表 2）．表 2 より，選択処理の確認ができた．誤りは 3 文あり，2 つに分類できた．以下に，多かった誤りの例を示す．

原文：トルコの「汗と絨毯」は、職人たちを追ったドキュメンタリーだ。

最適パターン：

$/[MT1]NP2(は\mid が)[、]/yf[MT3]NP4[MD5]$

選択型名：外延叙述型

抽出３つ組：上位:トルコの「汗と絨毯」, 下位:ドキュメンタリー, 追加情報:職人たちを追った

正解型名：範疇叙述型

正解３つ組：下位:トルコの「汗と絨毯」, 上位:ドキュメンタリー, 追加情報:職人たちを追った

この例では，範疇叙述型が選択され，下位語『トルコの「汗と絨毯」』，上位語「ドキュメンタリー」，および，追加情報「職人たちを追った」と抽出されれば成功である．しかし，型名の選択が誤りであった．誤り原因は，かぎ括弧内を固有名詞とみなす処理の不足である．

表 2　LOOCV における途中の意味解析の性能

選択処理	入力数	出力数	一致数	R	P	F 値
なし	37	572	37	1.00	0.06	0.11
1	37	133	37	1.00	0.28	0.44
1,2	37	85	37	1.00	0.44	0.61
1,2,3	37	78	37	1.00	0.47	0.64
1,2,3,4	37	63	37	1.00	0.59	0.74

6　おわりに

本稿では，名詞述語文の意味的な型の識別および意味情報の抽出のためのパターン集を作成した．結果として，37 文から 8 パターンおよび 37 通りの意味制約が得られた．実験により，パターン選択が良好であること，事例に基づく選択（手順５）は，効果があるが網羅性が不足することを確認した．

参考文献

[1] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦:“日本語語彙大系”, 岩波書店, 1997.

[2] 今田水穂:“日本語名詞述語文の類型と主語の意味分類について: 京都大学テキストコーパスと分類語彙表を用いた調査・検査”, 筑波大学文藝・言語学系, 文藝言語研究, 言語篇, 60, pp.25-48, 2011.

[3] 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.