名詞述語文の意味解析のためのパターン辞書の作成と運用

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
{s092055, tokuhisa, murakami, murata} @ ike.tottori-u.ac.jp

1 はじめに

本稿では，パターン辞書を用いた名詞述語文の意味解析の方法について議論する．この方法では，入力文をパターンと照合すると，複数の照合結果が得られる．その中から照合結果の選択を行うことで，意味が解析されたことになる．パターン辞書について，用言述語文を対象とするものとして日本語語彙大系 [1] が既に存在する．名詞述語文は今田の分類 [2] に基づく単文パターン集が作成されている [3]．ここで，照合結果選択の際，意味属性コードの近さおよび上下関係が利用されていた．しかし，重文複文については，主部や述部の具体性の判定において意味属性コードだけでは不足しており，主語の前の連体修飾表現の有無などの文構造の確認を加えて行う必要があるため，照合結果選択の条件が複雑になる．そこで，本稿では，照合結果選択の条件を判定するために機械学習を利用することを試みる．

以上に向けて，第 2 章では名詞述語文の分類を示す．第 3 章では，今田の例文，および，Wikipedia の文を対象に文の構造について分析を行う．特に分析における注意点を示す．第 4 章では分析結果をまとめてパターンの作成方法を示す．第 5 章では，入力文の意味を解析するためのパターン辞書の運用方法を示す．第 6 章ではパターンによる解析性能を実験により評価する．最後に第 7 章でまとめを述べる．

2 名詞述語文の分類

生産物名詞が主語となる名詞述語文は，主語と述語との意味関係に基づき，以下の 3 つに分類される [2].

属性叙述型

主語の価格や数量などの属性を述語で述べる文

例 1 このカレーは３００円だ。

外延叙述型

内包的概念を表す主語の外延を述語で述べる文

例 2 使った道具はドライバーだ。

範疇叙述型

主語の帰属する範疇を述語で述べる文

例 3 パソコンはただの道具だ。

3 名詞述語文の分析

3.1 目的

パターンの作成，および，運用を行う際，文の構造を把握するために，事例の分析を行う．

3.2 項目

今田の例文，および，Wikipedia の文を分析対象とする．Wikipedia は 2014 年 11 月 4 日時点のデータを使用する．本章では主に以下の点を明らかにする．

- 今田の 3 分類の是非
- 主語および述語についての抽出範囲
- はが構文の扱い
- 定形表現について

3.3 結果

3.3.1 今田の 3 分類の是非

複雑な文を分析していると，判断に迷うことがある．そこで，分類に対する考え方をまとめておく．

- 外延叙述型と属性叙述型は主語の外延を述べるか属性を述べるかの違いがあり別ものである．
- 範疇叙述型は属性叙述型に含まれるという指摘があるが，主語と述語の相対性を見ているので別のものと考える．
  - 属性叙述型は性質や程度などを表すもの，または，（拡大解釈すると）種類などを表すものなどがある．

    例 4（程度） このカレーは３００円だ。

    例 5（種類） 電子レンジは電化製品だ。

    例 4 では，述語は主語の価格を表しており，主語と述語に意味的な相対性はない．例 5 では，述語は主語の種類を表しており，主語と述語に包含関係があるため，主語と述語に意味的な相対性があるといえる．
  - 文中の主語や述語の抽象度は相対的に見るので，主語のみまたは述語のみから抽象度を確定させることはできない．
  - ゆえに，拡大解釈をせずに，例 5 は，範疇叙述型と考える．

以上より，本稿では，分類は今田の 3 分類とすることに問題ないと考える．

3.3.2 主語についての抽出範囲

文中の主部のうちどの範囲までをパターン処理上の主語として抽出すれば良いかについて考え方を示す．

例 6 初期の 自動車 は手作りのものである。

例 6′ 初期の自動車 は手作りのものである。

例 7 これを１本のケーブルで接続できるように端子を１つにまとめたもの がD端子である。

もし，主語の抽出範囲を例 6 の下線部のみとするならば，主語名詞だけでは具体的な実体ではなく，一般的なものを指すため，不適切である．これに対し，もし，述部と対応の取れる程度に具体的になるように例 6′ の下線部を主語の抽出範囲とするならば，主語の抽出範囲は連体修飾表現までとなる．しかし，例 6′ の基準を例 7 に用いると下線部のとおりの主語の抽出範囲となり，連体修飾節が接続詞や複数の動詞を含むという長い表現となるのだが，連体修飾節をパターンで全てをマッチさせても，型名の判定で全て使うとは限らないので，パターンで処理する上では不要である．したがって，「修飾表現が主語の前にあるか」という判断までを必要としつつ，例 6 の下線部を主語の抽出範囲と定める．

3.3.3 述語についての抽出範囲

主語と同様に，パターン処理上の述語としての抽出範囲の考え方を示す．

例 8 蒸気機関車は蒸気機関で動く 機関車 である。

述語の抽出範囲は，主語と同様の考えで，例 8 の下線部を述語の抽出範囲とし，述語の修飾語句を述語とは別の意味のある表現部分（後述の追加情報）と考える．

3.3.4 「はが構文」の扱い

「はが構文」は文の分析において判断を誤りやすい文であるので注意を述べる．

例 9 強度部材に用いられる材料は鋼鉄が主流である。

例 10* そのパスポートは私が落としたものだ。

「は格」の前の単語（例 9 では「材料」），および，「が格」の前の単語（例 9 では「鋼鉄」）の係り先が同一のものを「はが構文」という．例 10*のように「は格」の前の単語，および，「が格」の前の単語の係り先が別であれば，「はが構文」ではないことに注意を要する（図 1）．

一方，はが構文は（広い意味での）主語と述語の関係が 2 つ存在することにも注意を要する．

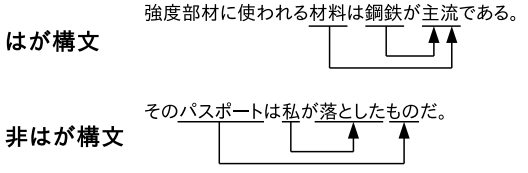

図 1 「はが構文」と「非はが構文」の例

3.3.5 定形表現について

定形表現を含む文についても判断を誤りやすいため注意を述べる．

例 11 ブルーチーズは、チーズの一種である。

「〜の一種」という表現がある．「の一種」は，その前の単語が主語の範疇を表すと判定する手がかりとなる表現である．これを本稿では定形表現と呼ぶことにする．定形表現を含む文に対しては，定形表現の前の単語に対して主語との関係を比較する．

他の定形表現を Wikipedia から収集を試みた．使用データは，「〜である」または「〜だ」の表現の前の単語を抽出し，頻度の高い単語から順に人手で分析を行った．上位 250 件までで，8 種類得た．定形表現を効率的に発見することは，今後の課題とする．

得た表現 一種，１種，１つ，一つ，ひとつ，仲間，総称，一分野

4 パターンの作成

4.1 パターン化基準

パターンの作成にあたって，3 章の調査を基にパターン化基準を作成する．これにより，パターン作成者個人に依存する揺れを抑える．全ての型名で共通する基準と，各型名での基準があるので，共通 (G)，属性叙述型 (A)，外延叙述型 (E)，範疇叙述型 (C) の順で記述する (表 1)．パターン化基準は，作業者に対する指示であり，「確認する」こと，「記述する」こと，および，「注意する」ことが混在している．

パターン作成の成果物はパターン辞書である．パターン辞書は多数のエントリで構成する．1 つのエントリは，「原文（1 文以上）」，「パターン」，「選択条件」，および，「応用情報」で構成される．

原文は，パターン，選択条件，および，応用情報を作成する際に参照する文である．

パターンは，字面，変数，および，記号で構成される．字面は，パターンに文字を記載したもので，照合の際は，字面部分は完全一致のみ可能である．変数は，品詞情報を基に単語，句，節を表す変数である．N は名詞，MT は連体修飾表現，MD は判定詞を含むモダリティ表現を表す変数である．記号は，パターンに汎用性を持たせるために使用する．「/」は，任意の単語列に適合する記号であり「離散記号」と呼ぶ．「|」は，記号の前後の字面や変数を選言的に照合する．大括弧は，括弧内の字面や変数の有無を任意化して照合する．

選択条件は，照合結果の選択を行う際に使用する条件で，「修飾表現が主語の前にあるか」というフラグである（3.3.2 節参照）．また，変数の意味属性制約もここに該当する．

応用情報とは，パターンを照合した際，出力される情報であり，意味解析結果として得たい情報のテンプレートである．名詞述語文の場合，名詞述語文の叙述の型名，および，3 つ組がある．3 つ組の情報は，上位語，下位語，追加情報，実体，属性，および，属性値であり，この組み合わせを 3 つ組とする．追加情報は，述語の前にある連体修飾表現である（3.3.3 節参照）．

4.2 作成結果

文献 [2] に示された例文 33 文，および，2014 年 11 月 4 日時点の Wikipedia「自動車」ページから得た具体物が

表 1 パターン化基準

基準 ID	パターン化基準
G1	はが構文と非はが構文の 2 種類が存在する.
G2	主語はパターンでは名詞（連続した名詞の場合一つの名詞とみなす）と前後の接辞まで含め変数化する.
G3	主語の前に連体修飾表現（「A の B」型名詞句の場合は「A の」をこの表現に含める）がある場合，変数化し，選択条件にフラグ T を記す.
G4	主語の前に連体修飾表現がない場合，選択条件にフラグ F を記す.
G5	はが構文の場合，助詞は，1 つ目は「は」で 2 つ目が「が」で，助詞の直後の読点まで字面に残す.
G6	非はが構文の場合，助詞は，「は」または「が」で，助詞の直後の読点まで字面に残す．また，「は」および「が」の選択化，読点の任意化を行う.
G7	述語はパターンでは名詞（連続した名詞の場合一つの名詞とみなす）と前後の接辞まで含め変数化する.
G8	述語の前に連体修飾表現がある場合，それが追加情報で，変数化し，そうでない場合，追加情報は付けない．また，助詞と追加情報の間の単語を削除し，離散記号を記す.
G9	文末の判定詞を変数化する.
G10	主語または述語の具体性が低い名詞と判断した場合は，該当名詞を字面に残したパターンを別に作成し，計 2 パターン用意する.
G11	型名を確認し，応用情報に記す.
A1	属性叙述型から取り出される情報は，実体，属性，属性値の 3 つであり，型名および 3 つ組を応用情報とする.
A2	主語は実体を表す.
A3	追加情報および述語は事例によってそれぞれが属性または属性値になり得る.
A4	取り出される情報の述語は，抽象的な意味を持つ単語である.
E1	外延叙述型から取り出される情報は，上位語，下位語，追加情報の 3 つであり，型名および 3 つ組を応用情報とする.
E2	主語は上位語を表す.
E3	述語は下位語を表す.
E4	相対的に見て主語が上位語で，述語が下位語である.
E5	述語が固有名詞で，主語は固有名詞でない場合は外延叙述型といえる.
E6	主語が形式名詞のように抽象度が高い場合は，外延叙述型といえる.
C1	範疇叙述型から取り出される情報は，下位語，上位語，追加情報の 3 つであり，型名および 3 つ組を応用情報とする.
C2	主語は下位語を表す.
C3	述語は上位語を表す.
C4	相対的に見て主語が下位語で，述語が上位語である.
C5	主語が固有名詞で，述語は固有名詞でない場合は範疇叙述型といえる.
C6	主語の前に指示詞がある場合は，主語の抽象度が低いとみなすことができる.
C7	判定詞の前に単語自体で文を範疇叙述型と判定できる定形表現がある場合は，該当部分を字面に残し，直前の名詞と接辞までを述語のように扱う.

主語の名詞述語文 8 文の計 41 文を基にパターンを作成したところ，21 エントリを得た．以下にエントリの作成例を示す．なお，本稿では，名詞述語に対する主語が存在し，かつ，末尾に判定詞が存在する名詞述語文を対象とする.

エントリ 1.
原文 1: カレーは三百円だ。
原文 2: タンクはグラスファイバー製だ。
原文 3: 実験器具は実費だ。
パターン: $/N1$(は | が)[、]$/N2\ MD3$
選択条件:
主語前: T
応用情報:
型名: 属性叙述型
3 つ組: （実体:$N1$, 属性:$N2$, 属性値:$N2$）

エントリ 2.
原文 4: 使った道具はドライバーだ。
原文 5: ともに使われた凶器は、けん銃だった。
パターン: $/MT1\ N2$(は | が)[、]$/N3\ MD4$
選択条件:
主語前: T
応用情報:
型名: 外延叙述型
3 つ組: （上位語:$N2$, 下位語:$N3$, 追加情報:nil）

エントリ 3.
原文 6: パソコンはただの道具だ。
原文 7: 抗がん剤が自社の製品だった。
原文 8:「しめ縄」は、悪霊の侵入を防ぐ道具だった。
パターン: $/N1$(は | が)[、]$/MT2\ N3\ MD4$
選択条件:
主語前: F
応用情報:
型名: 範疇叙述型
3 つ組:（下位語:$N1$, 上位語:$N3$, 追加情報:$MT2$）

5 パターンの運用

まず，入力文を全てのパターンと照合すると，照合結果が複数得られる．照合結果には応用情報および変数へのバインド値が含まれている．1 つのパターン（エントリ）においても，バインドが複数通りになりえるので，照合結果は複数通りになる．そこで，次に，照合結果の選択を行う．表 1 のパターン化基準を参考に作成した素性生成ルール（表 2）で照合結果毎に真理値タプルを作成する（図 2）．このタプルをベクトル化し，機械学習で各照合結果を採用すべきか破棄すべきか識別を行う．最後に，採用した照合結果の応用情報（バインド済み）を，入力文の意味解析結果として出力する.

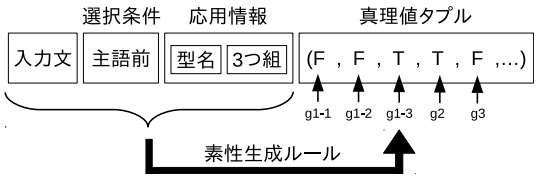

図 2 真理値タプルのイメージ

6 実験

6.1 クローズドテスト

今田の例文（33 文，正解データ 33 件），2014 年 11 月 4 日時点の Wikipedia の「自動車」ページの名詞述語文（8 文, 正解データ 11 件）それぞれに対し，leave-one-out cross-validation を行った．比較手法は，意味属性コードの近さおよび上下関係を使用する従来手法である [3]．また，一致数は，型名および 3 つ組が正解データと全て同一の場合にカウントする．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 3,

表 2 素性生成ルール

素性 ID	素性生成ルール (T/F)
g1-1	パターンの原文は，はが構文か
g1-2	マッチした文は，はが構文か
g1-3	g1-1 および g1-2 の判定は一致しているか
g2	主語はあるか
g3	パターン原文の主語の前に連体修飾節があるか
g4-1	マッチした文の主語の前に連体修飾節があるか
g4-2	g3 および g4-2 の判定は一致しているか
g5	はが構文かつ，助詞「は」および「が」があるか
g6	非はが構文かつ，助詞「は」または「が」があるか
g7	述語はあるか
g8-1	パターン原文に追加情報はあるか
g8-2	マッチした文に追加情報はあるか
g8-3	g8-1 および g8-2 の判定は一致しているか
g9	文末に判定詞はあるか
g10	マッチしたパターンに名詞の字面が残っているか
g11-1	照合結果で想定されている型名は何か（非真理値）
a4-1	属性の意味属性コードは全て抽象か
a4-2	属性の意味属性コードは全て具体か
a4-3	属性の接尾辞は “数え方の辞典”[4] にあるか
a4-4	属性の品詞は用言性名詞か
a4-5	属性の品詞は形容動詞転成型名詞か
ec4-1	意味属性コード上で主語は述語の上位か（不明は F）
ec4-2	意味属性コード上で主語は述語の下位か（不明は F）
ec4-3	主語と述語の共通する意味属性コードはあるか
ec4-4	主語に意味属性コードはあるか
ec4-5	述語に意味属性コードはあるか
e5-1	述語は固有名詞か
e5-2	述語はカギ括弧で囲われているか
e5-3	述語はアルファベットの単語か
e6	主語は形式名詞か
e7	主語の前に連体修飾表現があり，追加情報がないか
c5-1	主語は固有名詞か
c5-2	主語はカギ括弧で囲われているか
c5-3	主語はアルファベットの単語か
c6	主語の直前に指示詞があるか
c7	マッチしたパターンに定形表現があるか

4)．ここで，$R = \langle$ 一致数 $\rangle/\langle$ 採用すべき数 $\rangle$, $P = \langle$ 一致数 $\rangle/\langle$ 採用数 $\rangle$, F 値 $= 2PR/(P+R)$ である．$T = \langle$ フィルタリングの対象数 $\rangle$ であるが，照合結果数であるので，入力文数（33 や 11）よりも多い．

表 3 今田の例文の意味解析の性能（クローズド）

手法	T	再現率 R	適合率 P	F 値
従来手法	98	0.91 (30/33)	0.63 (30/48)	0.74
提案手法	98	0.70 (23/33)	0.79 (23/29)	0.74

表 4 「自動車」文の意味解析の性能（クローズド）

手法	T	再現率 R	適合率 P	F 値
従来手法	34	0.45 (5/11)	0.36 (5/14)	0.40
提案手法	34	0.36 (4/11)	0.57 (4/ 7)	0.44

6.2 オープンテスト

2014 年 11 月 4 日時点の Wikipedia の「潜水艦」ページの名詞述語文（26 文，正解データ 27 件）のオープンテストを行った．学習データは，6.1 節の 2 種類のデータを両方使用した（132 件）．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 5）．

表 5 「潜水艦」文の意味解析の性能（オープン）

手法	T	再現率 R	適合率 P	F 値
従来手法	127	0.59 (16/27)	0.55 (16/29)	0.57
提案手法	127	0.56 (15/27)	0.60 (15/25)	0.58

6.3 考察

表 3, 4, 5 全てで，適合率 P は改善されている．表 4, 5 については，F 値もわずかに改善されている．しかし，再現率 R は減少した．以下に一致の例を示す．

入力文: 攻撃型潜水艦は、魚雷や機雷などを主兵装とし、敵の水上艦艇や潜水艦などの攻撃を任務とする潜水艦である。
選択パターン: $/N1(\text{は} \mid \text{が})[\text{、}]/MT2\ N3\ MD4$
選択型名: 範疇叙述型
抽出 3 つ組: （下位語:攻撃型潜水艦, 上位語:潜水艦, 追加情報:敵の水上艦艇や潜水艦などの攻撃を任務とする）
正解型名: 範疇叙述型
正解 3 つ組: （下位語:攻撃型潜水艦, 上位語:潜水艦, 追加情報:敵の水上艦艇や潜水艦などの攻撃を任務とする）

この例では，型名および 3 つ組が一致しており，正解であることを確認できる．次に誤りの例を示す．

入力文: 当初から物資運搬を想定して建造された最初の輸送型潜水艦は、第一次大戦期のＵ１５１型Ｕボートである。
選択パターン: $/MT1\ N2(\text{は} \mid \text{が})[\text{、}]/MT3\ N4\ MD5$
選択型名: 範疇叙述型
抽出 3 つ組: （下位語:輸送型潜水艦, 上位語:Ｕ１５１型Ｕボート, 追加情報:第一次大戦期の）
正解型名: 外延叙述型
正解 3 つ組: （上位語:輸送型潜水艦, 下位語:Ｕ１５１型Ｕボート, 追加情報:第一次大戦期の）

この例では，選択された型名および 3 つ組の組み合わせが誤っていた．誤り原因は，述語の「Ｕ１５１型Ｕボート」を固有名詞として処理できなかったことである．

7 まとめ

本稿では，名詞述語文の意味解析のためのパターン辞書の作成方法を示し，運用方法の中でも特に，照合結果選択の条件を判定するために機械学習を利用する方法を示した．実験により，従来手法よりも性能がわずかに改善されたことを確認した．

今後の課題は，本稿では取り扱わなかった主語なし文や末尾に判定詞のない名詞述語文の対応，および，重文複文全般を網羅している鳥バンク [5] と本稿で作成したパターン辞書との比較である．

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] 藤原竜樹, 徳久雅人, 村上仁一, 村田真樹:“意味類型化のための名詞述語文のパターン化”, 情報処理学会第 76 回全国大会講演論文集, 2, pp.157-158, 2014.
[4] 飯田朝子, 町田健:“数え方の辞典”, 小学館, 2004.
[5] 鳥バンク : http://unicorn.ike.tottori-u.ac.jp/toribank/