文末述語における機能表現検出と文間接続関係推定への応用

佐野 正裕　　佐藤 理史　　宮田 玲
名古屋大学大学院工学研究科
sano.masahiro@nagoya-u.jp

1 はじめに

日本語の文末部分には、モダリティやテンスといった、文法的機能を担う表現がよく現れる。たとえば、文 (1a) の「〜てみよう」は提案（モダリティ）を、文 (1b) の「〜なければならない」は義務（モダリティ）を表す機能表現である。さらに、文 (1b) において、仮定を意味する「〜とする」は、接続詞「たとえば」と呼応しやすい機能表現の一つである。

(1) a. 実際に即して考えてみよう。
b. **たとえば**学校で宿題が出て、次の授業までにある文章なり数式をおぼえていかなければならない とする。
(出典: 早稲田大学入試 [1])

文末の機能表現に着目した研究には、文末の助動詞や表明思考動詞を利用し、ウェブページが主観情報中心か客観情報中心かの分類に試みた例 [2] がある。さらに、特定パターン（構文パターン）によって、文から内容に関わる語や文節を削除してスコアを計算し、二文間の接続関係推定を行った例 [3] もある。

しかし、一般に、文末のどの範囲を文末述語として認定すべきかは、それほど自明ではない。たとえば文 (2) では、我々は直感的に「見つかるわけはありません」を述語部分だと考えるが、KNP はこの部分を 3 文節（ | が文節境界）に認定する。

(2) けれども、| 見つかる | わけは | ありません。
(出典: 宮崎大学入試 [1])

さらに、文末述語をどのように構成要素に分解するかという点も検討が必要である。文 (2) の「わけはありません」の中核は「わけがない」（不可能）であり、これに取立助詞「は」と丁寧体が加わった形式と解釈できる。しかしながら、既存の解析器では、「わけがない」の存在を自然には特定できない。以上のような、解析結果が我々の直感と合わない原因はともに、解析器における述語複合辞の認定が不十分な点にある。

本稿では、述語複合辞に着目することで、我々の直感に合った、文末述語範囲の同定と構成要素への分解（機能表現検出）を行う「文末述語解析システム」Panzer について述べる。さらに、Panzer の有効性を確かめるために取り組んだ、文末述語を用いた二文間の接続関係推定について述べる。

2 文末述語解析システム

2.1 解析方針

文末述語解析システム **Panzer** (Predicate Analyzer) は、入力された日本語文の文末述語範囲を同定し、構成要素に分解する。一般に文末述語は、核となる要素（おおむね、動詞、形容詞、名詞）である**内容部**と、助詞や助動詞などの機能語列である**機能部**によって構成されると考えることが多い。我々もこの考え方に則るとともに、次の 2 つの方針を取り入れる。

1. 機能語として認定する要素を厳格に定義する。
2. 認定する機能語に、述語複合辞を含める。
ここで、文末に頻出する述語複合辞は、可能な限り網羅的に辞書登録しておく。

解析にあたっては、KNP の最終文節を、暫定的な文末述語とみなす。文末側の形態素からさかのぼって機能語（述語複合辞を含む）を一つずつ認定することで、構成要素に分解する。さらに、述語複合辞が文節境界をまたぐ場合は、その前の文節を繰り入れて文末述語とする。これにより、文末述語範囲の同定と、構成要素への分解を実現する。

2.2 解析例

文 (1a), (1b), (2) に対する Panzer の解析結果を、図 1 に示す。各結果の 1 行目は、JUMAN+KNP による解析結果（_は形態素境界、| は文節境界）である。2 行目の [·] の部分が、文末述語解析の結果を示す。そのうち、{ · }の部分は内容部、残りは

(1a) 実際_に | 即して | 考えて_みよう_。
→ [{ 考える }, ⟨ テ助動詞みる ⟩, ⟨ 意志形 ⟩, ⟨ 句点 ⟩] 【動詞＋助動詞, 動作, 普通体】

(1b) たとえば | 学校_で | 宿題_が | 出て_、 | 次の | 授業_まで_に | ある | 文章_なり | 数式_を | おぼえて_いか_なければ_なら_ない_と_する_。
→ [{ おぼえる }, ⟨ テ助動詞いく ⟩, ⟨ なければならない ⟩, ⟨ とする ⟩, ⟨ 句点 ⟩]
【動詞＋助動詞, 動作, 普通体, 義務 (なければならない), 判定/仮定 (とする)】

(2) けれども_、 | 見つかる | わけ_は | あり_ませ_ん_。
→ [{ 見つかる }, ⟨ わけがない ⟩, ⟨ 格取立は ⟩, ⟨ 接尾辞ます ⟩, ⟨ 句点 ⟩]
【動詞＋助動詞, 動作, 丁寧体, 不可能 (わけがない)】

図 1: 文 (1a), (1b), (2) の Panzer による解析結果

機能部（⟨ 機能語 ⟩ の列）である。ここで、それぞれの機能語（⟨ テ助動詞みる ⟩, ⟨ なければならない ⟩ など）は、日本語文生成器 HaoriBricks3 [4] の Brick に対応しており、HaoriBricks3 を用いて元の文を復元することが可能である。

2.3 述語複合辞

述語複合辞は、複数の形態素から構成されているにもかかわらず、全体として一つの助動詞のように働く表現である。たとえば、「かも_しれ_ない」は、JUMAN は 3 形態素に認定するが、全体として、可能性を表す表現だと考えるのが自然である。

一般に、述語複合辞の直前にある述語の形式（活用型と活用形）は、限定される。たとえば、「〜ずにはいられない」の直前は述語の未然形に、「〜かもしれない」の直前は述語の終止形・タ形（述語が名詞の場合、判定詞「だ」の終止形は消失する）に限られる。

Panzer では、述語複合辞を構成する形態素と、その直前にある述語の形式を含めた、形態素列照合パターン（述語複合辞の認定ルール）を、具体例から自動生成することによって、述語複合辞の認定を実現する。認定すべき述語複合辞として、文献 [5] を参考にし、239 種類（457 エントリ）を実装した。

2.4 解析手順

文 (2) を例に、Panzer による解析手順を説明する。

1. **前処理** JUMAN+KNP による解析を行い、暫定的な文末述語範囲を同定する。あわせて、次の丁寧体を正規化する。
   - あり_ませ_ん → ない_ます
   - あり_ません_でした → ない_ました
   - (書き)_ませ_ん → (書か)_ない_ます
   - (書き)_ませ_ん_でした → (書か)_ない_ました

   これは、要素 ⟨ ない ⟩ を末尾に含む述語複合辞（「わけはない」と「わけはありません」など）の認定ルール (2.3 節) を共通化し、ルール数を減らすためである。
2. **構成要素の認定** 文末の形態素から順に、次に合致する部分を抽出・認定する。
   (a) 述語複合辞の認定ルールに合致すれば、その構成形態素を、述語複合辞を構成する機能語 (Brick) 列に変換する。このとき、述語複合辞が文節境界をまたぐ場合は、前の文節を繰り入れる。
   (例) わけ_は_ない
   → ⟨ わけがない ⟩ + ⟨ 格取立は ⟩
   (b) 辞書に定義した機能語もしくは機能標識であれば、Brick に変換する。
   (例) ⟨ 句点 ⟩, ⟨ 接尾辞ます ⟩
   (c) いずれにも当てはまらない場合、内容語と認定する。
3. **内容語の結合** 内容語が見つかった場合、その後、最初に到達した文節境界までの形態素を、内容語として結合する。
4. **不要な Brick の削除** 元の文への復元に不要な活用形を示す Brick は、出力する結果から削除する。

2.5 文末述語に対する属性付与

文の基本的情報として、Panzer の出力する文末述語の情報から、次の属性を付与する（図 1 の【・】部分）。

- **述語形式** 内容語の品詞 ＋ （助動詞相当の機能語の有無） ＋ （終助詞の有無）

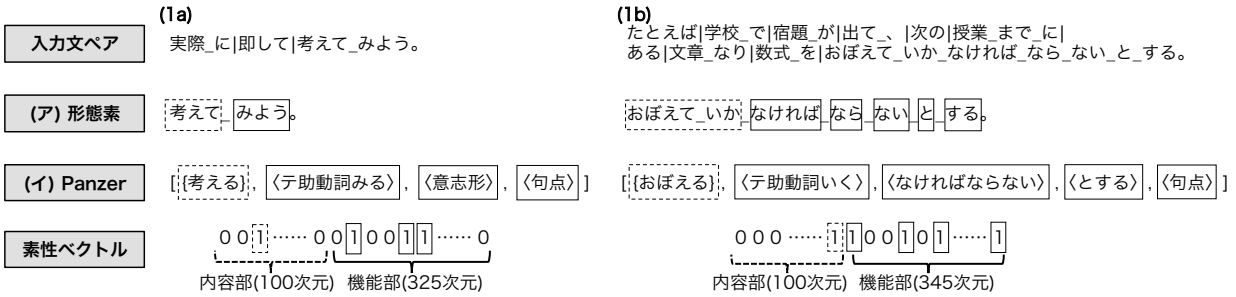

図 2: 文 (1a), (1b) の素性ベクトルへの変換手順

- **述語タイプ**　「動作」もしくは「状態」（次の Brick を含むか、内容語が動詞以外の場合）
〈テ助動詞いる〉, 〈取立助動詞ある〉, {いる}, {ある}, {できる}, {要る}, {異なる}, {違う}
- **常敬体**　「普通体」もしくは「丁寧体」（〈デス体〉, 〈マス体〉, 〈接尾辞です〉, 〈接尾辞ます〉を含む場合）
- **テンス**　〈タ〉の有無
- **アスペクト・モダリティ**　アスペクト・モダリティとして働く述語複合辞とその大まかな意味一覧

3　文末述語を用いた文間接続関係推定

文間の接続関係は、文章を解析する大きな鍵となる。特に文末述語は、文 (1b) の「〜とする」のように、接続関係推定に大きく寄与する要素を含むことが期待される。本節では、文末述語を用いて、文間の接続関係をどれだけ推定できるのか、実験的に調査した結果を述べる。

3.1　問題設定

文 (1a), (1b) のように連続した二文の文末述語を与えて、その二文をつなぐ接続表現を推定する問題として考える。推定のために多項ナイーブベイズ学習器を構築した。その素性には、内容語（原形）と機能語の存在に基づくワンホット表現のベクトルを用いた (図 2)。内容語は、対象データに出現する全内容語のうち、出現頻度が上位 99 位まではリテラルとして変換し、100 位以下の内容語は、その他として一つにまとめた[1]。

[1] これ以外に、次の 3 種類の設定も試した。本文に示したものは、最も結果が良かった設定である。

(1) 多項ナイーブベイズの素性を、以下によって作成するもの。
  (a) 内容語を用いず、機能語のワンホット表現のみを用いる。
  (b) 内容語に、Word2Vec (100 次元ベクトル) を用いる。

(2) あらかじめ対象データ中の機能語の出現頻度を求めておき、出現頻度の相乗平均に基づくスコアを用いて判定するもの。

表 1: 機能語の内訳（種類）

述語複合辞	174
助動詞	62
助詞	73
接尾辞	21
活用形	27
その他（判定詞、取立など）	11
その他（記号など）	13
合計	381

Panzer の有効性を確かめるため、次の 2 つの手法を比較した。

- **手法 (ア) 形態素ベース**　KNP の最終文節の形態素を用いるもの。内容部と機能部の分割点は、その文節の最も左にある助詞または助動詞の前とした。ただし、分割点の左の、動詞または接尾辞の「いる」「ある」「する」「なる」「できる」「ない」は、機能部とした。
- **手法 (イ) Panzer**　Panzer の出力を用いるもの。

3.2　対象データ

対象データとして、対象の接続表現を文頭に持つ文と、その直前の文のペアを収集した。対象とした接続表現は、「だから」「なぜなら」「つまり」「たとえば」「さらに」「しかし」の 6 種類である。接続表現ごとに、次のデータを収集して利用した。

- **訓練データ**　BCCWJ [6] から 5000 件
- **テストデータ**　大学入試国語現代文で用いられた評論文 [1] から 100 件

このデータに Panzer を適用したところ、文末述語を構成する Brick の総出現数は、内容語 11,180 種類、機能語 381 種類であった。機能語は、大きく分けて、表 1 のように整理できる。また、文末述語に含まれる機能語の平均個数は、一文あたり 2.86 であった。

表 2: 手法 (ア) 形態素ベース の混同行列

正解＼出力	だから	なぜなら	つまり	たとえば	さらに	しかし	Recall	F 値
だから	16	6	19	6	28	25	16/100 (16.0%)	20.9%
なぜなら	4	78	6	0	6	6	78/100 (78.0%)	71.9%
つまり	7	8	37	3	30	15	37/100 (37.0%)	40.7%
たとえば	6	9	5	21	42	17	21/100 (21.0%)	30.0%
さらに	10	11	7	4	49	19	49/100 (49.0%)	34.1%
しかし	10	5	8	6	32	39	39/100 (39.0%)	35.3%
Precision	16/53 (30.2%)	78/117 (66.7%)	37/82 (45.1%)	21/40 (52.5%)	49/187 (26.2%)	39/121 (32.2%)	240/600 (40.0%)	F 値平均: 38.8%

表 3: 手法 (イ) Panzer の混同行列

正解＼出力	だから	なぜなら	つまり	たとえば	さらに	しかし	Recall	F 値
だから	19	8	21	10	22	20	19/100 (19.0%)	25.5%
なぜなら	6	82	4	2	2	4	82/100 (82.0%)	74.5%
つまり	4	9	43	8	19	17	43/100 (43.0%)	44.5%
たとえば	13	5	8	33	22	19	33/100 (33.0%)	35.9%
さらに	3	7	6	22	46	16	46/100 (46.0%)	38.6%
しかし	4	9	11	9	27	40	40/100 (40.0%)	37.0%
Precision	19/49 (38.8%)	82/120 (68.3%)	43/93 (46.2%)	33/84 (39.3%)	46/138 (33.3%)	40/116 (34.5%)	263/600 (43.8%)	F 値平均: 42.7%

3.3 実験結果と考察

手法 (ア), (イ) のテスト結果をそれぞれ、表 2, 表 3 に示す。なお、F 値平均は、接続表現ごとに求めた F 値（Recall と Precision の調和平均）の平均値を示す。

形態素ベースによる手法 (ア) では 600 文中 240 文 (40.0%) に対して、Panzer による手法 (イ) では 600 文中 263 文 (43.8%) に対して、正しい接続表現を推定できた。いずれの手法でも、接続表現によって正答率には差がある。「なぜなら」は、他の接続表現に比べて高い (82%) Recall であった。このことは、「なぜなら」は「からである」のような典型的な文末機能表現によって推定できるという直感にも整合する。一方、「だから」は、ほぼチャンスレベル (19%) の Recall であった。このような Recall の低い接続表現の推定には、二文の内容語同士の同意・反意関係や、文末述語以外の部分、一文以上前後の文の情報が必要であると考えられる。

F 値は、すべての接続表現についておおよそ 2〜6% 程度向上している。さらに、手法 (ア), (イ) に対して McNemar 検定を適用したところ (表 4)、有意水準 5%において統計的に有意な差が見られた。このことから、接続関係を推定するために有用な情報が、Panzer による機能語・述語複合辞の認定によって検出できていることが示唆される。

表 4: McNemar 検定の行列

(ア) 形態素ベース	(イ) Panzer 正解	(イ) Panzer 不正解	計
正解	193	47	240
不正解	70	290	360
計	263	337	600

謝辞 本研究は JSPS 科学研究費基盤研究 (B)「日本語文章の構造モデルとその段階的詳細化による文章自動生成機構」(課題番号 18H03285) の助成を受けている。

参考文献

[1] 現代文問題データベース Vol.6. 明治書院, 2017.

[2] 松本章代, 小西達裕, 高木朗, 小山照夫, 三宅芳雄, 伊東幸宏. 文末表現を利用したウェブページの主観・客観度の判定. 第 1 回データ工学と情報マネジメントに関するフォーラム (DEIM), A5-4, 2009.

[3] 齋藤真実, 山本和英, 関根聡. 大規模テキストを用いた 2 文間接続関係の同定. 言語処理学会第 12 回年次大会発表論文集, pp. 969–972, 2006.

[4] 佐藤理史. HaoriBricks: ブロック玩具に学ぶ日本語文章生成ライブラリ. 言語処理学会第 23 回年次大会発表論文集, pp. 20–23, 2017.

[5] 近藤泰弘, 坂野収, 多田知子, 岡田純子, 山元啓史. BCCWJ 複合辞辞書, 2011. http://japanese.gr.jp/documents/data/bccwjhukug-ouprint.pdf (2020 年 1 月 6 日確認).

[6] 山崎誠（編）. 書き言葉コーパス —設計と構築—. 朝倉書店, 2014.