文末述語解析と文末変換

Sentence-End Predicate Analysis and Transformation

名古屋大学大学院工学研究科教授
佐藤　理史

京都大学大学院工学研究科電気工学第二専攻博士課程研究指導認定退学。博士（工学）。北陸先端科学技術大学院大学、京都大学を経て、2005 年より現職。

1 はじめに

日本語の文末には、述語があるのが普通であるが、どの範囲を文末述語とみなすべきかは、それほど自明ではない。たとえば、以下の文では、どの範囲を文末述語とみなすべきであろうか。

（1）実際に即して考えてみよう
（2）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
（3）けれども、見つかるわけもありません

徹頭徹尾、文末述語を形式的に考えるのであれば、次のようにみなすのが標準的であろう。

（1）みよう
（2）する
（3）ありません

しかしながら、文意の把握を目的とするならば、文末述語を次のように捉える必要があるだろう。

（1）考えてみよう
（2）おぼえていかなければならないとする
（3）見つかるわけもありません

現在の日本語の文解析の標準的な方法は、文節係り受け解析である。その出力である文節係り受け構造では、暗黙的に最終文節（係り受け構造の根となる文節）が文末述語とみなされる。たとえば、日本語解析システム Juman/KNP は、これらの文を以下のように解析する。ここで '_' は形態素境界、'|' は文節境界を表す（これ以外に、品詞や活用型・活用形に関する情報を出力するが、ここでは省略した）。

（1）実際 _ に | 即して | 考えて _ みよう
（2）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
（3）けれども _、| 見つかる | わけ _ も | あり _ ませ _ ん

ここに示したように、（1）（2）に対しては、望ましい部分が最終文節（＝文末述語）として認定される。しかしながら、（3）は「ありません」が最終文節となる。

言語処理の応用システムでは、解析システムの後段に、なんらかの意味的な処理が接続するのが普通である。そのため、解析システムは、意味を担う単位を認定できることが望ましい。上記に示したように、Juman/KNP は、その方向を志向しているようである。

しかし、次のような課題が存在する。

- 複合辞「なければならない（義務）」は、明示的には認定されない。
- 「見つかるわけもありません」が３文節と認定される。「わけもありません」は、複合辞「わけがない（不可能）」の「が」が「も」で取り立てられて「わけもない」となり、さらに、これが丁寧の意味が付与されて、「わけもありません」となっていることが認定されない。

これらの課題は、述語に接続する機能要素（助動詞、複合辞など）としてどのようなものを設定するかを厳格に定め、それらを自動認定することなしには解決しない。特に、複合辞（複数の形態素から構成されるが、全体として一つの意味・機能を担う機能語相当表現）の認定は不可欠である。この例の場合、

● 「なければならない（義務）」という単位を設定・認定すべきである。
● 「わけもありません」を、『「わけがない（不可能）」＋取立助詞「も」＋丁寧体』と分解すべきである。

我々は、文末述語の範囲を同定すると同時に、文末述語を適切な構成要素に分解する処理が、いくつかの応用において必要と考え、これを文末述語解析と名付けた[1]。本稿では、これを実現する文末述語解析器 Panzer について述べる。

Panzer は、文を合成するためのドメイン特化言語である HaoriBricks3（HB3）[2] と不可分の関係にある。この２つのシステムを組み合わせることによって、文の文末形式を加工すること（文末変換）が可能となる。

2 文末述語解析器 Panzer

文末述語解析器 Panzer の概要を図１に示す。Panzer は、文字列の文を受け取り、前処理として、Juman による形態素解析と KNP による文節解析（文節認定）を実行したのち、文末述語の範囲を決定し、その部分を構成要素の列として出力する。図１の例文では、KNP の文節解析の出力の最後の３文節が、文末述語として認定される。

Panzer が文末述語として認定するのは、内容語と、それに続く機能語・機能要素の列である。図１の例文では、内容語「書く」と、それに続く４つの機能語・機能要素を認定している。「テ助動詞いる」は、動詞のテ形に後続する「いる」で、状態（アスペクト）を表す。「テンス有標」はテンス過去が付与されていることを表す。「にちがいない」は直感的確信（モダリティ）を表す複合辞である。「敬体ます」は丁寧体（「ます」と「です」の両方が可能な場合は「ます」を選択する）であることを表す。このように、Panzer の出力する構成要素は、述語の意味の決定に必要な文法的機能（テンス、アスペクト、モダリティ等）を担う単位である。

Panzer が認定する機能語・機能要素は、おおよそ、つぎのように分類できる。

（1）機能語（助詞、助動詞、述語に関連する接尾辞）
（2）複合辞
（3）文法機能要素（テンス、敬体、活用形など）

これらの要素は、列挙という形で完全に定義され、それぞれに固有の ID が付与されている。Panzer の出力は、内容語を除きすべてこの ID で表現される。

Panzer が行う処理は、それほど複雑ではない。形態素解析列としての文の末尾から長さ n の形態素列を切り出し、それを対応する ID に変換することを繰り返すだけである。多くの場合は、長さ n=1 であるが、複合辞を認定する場合は１より長くなる。内容語が見つかると、そこから一番近い文節境界までを内容語（複合語）として切り出し、処理を終了する。この処理は、述語の核となる内容語が、かならずしも１形態素として認定されるわけではないことに対処するためである。たとえば、複合動詞（「書き続ける」）は２形態素として認定されるのが普通である。図２に、Panzer の実行例を示す。

Panzer の一番の特徴は、文末述語の認定結果として出力される構成要素（内容語、機能語・機能要素）が、HB3 のブリック（文を組み立てる部品）に直接、対応している点にある。内容語の部分は、内容語を作り出すブリックとして出力され、ID として出力される機能語・機能要素は、その ID がそのままブリック名となる。こ

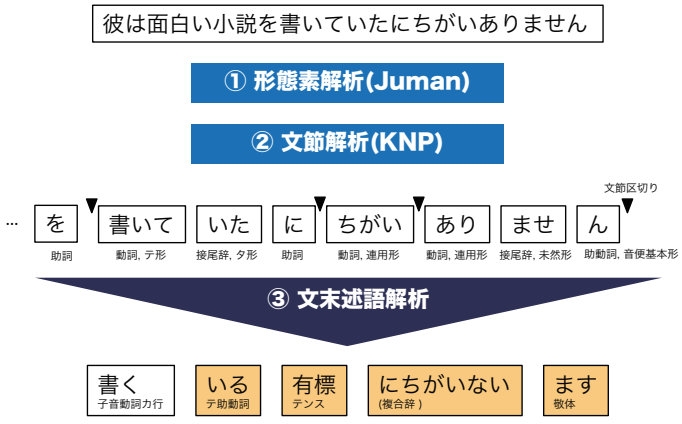

図１　Panzer の概要
形態素解析と文節解析の結果を利用して、文末述語の範囲を同定し、それを構成要素の列として出力する

彼は、面白い小説を書いていたにちがいありません
[{:lex=>"書く", :ctype=>:子音動詞型カ行}, :テ助動詞いる, :テンス有標, :にちがいない, :敬体ます]

実際に即して考えてみよう
[{:lex=>"考える", :ctype=>:母音動詞型}, :テ助動詞みる, :意志形]

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
[{:lex=>"おぼえる", :ctype=>:母音動詞型}, :テ助動詞いく, :なければならない, :とする]

けれども、見つかるわけもありません
[{:lex=>"見つかる", :ctype=>:子音動詞型ラ行}, :わけがない, :格取立も, :敬体ます]

小説を書き続けることもできたかもしれないのです
[{:lex=>"書き続ける", :ctype=>:母音動詞型}, :ことができる, :格取立も, :テンス有標, :かもしれない, :連体助動詞のだ, :敬体です]

図２　Panzer の実行例
それぞれの文の次の行が、Panzer の解析出力である。{} で括られた部分が内容語を表し、それ以外は、機能語・機能要素の ID を表す。

の特徴のおかげで、HB3 を使うと、Panzer の出力から文末述語の表層文字列を完全に復元することができる。このことは、Panzer の出力には、文末述語の表層形を決定するのに必要十分な情報が含まれていることを意味する。

Panzer の出力を見ると、情報が足りないのではないかと思われるかもしれない。たとえば、図２の最初の例の内容語「書く」には活用型の情報（「子音動詞型カ行」）は含まれるが、活用形の情報は含まれない。しかしながら、その活用形は、その直後が「テ助動詞いる」であることから、テ形になることが自動的に定まる。「テ助動詞いる」の活用形は、直後が「テンス有標」であることからタ形と定まる。複合辞「にちがいない」は、直後の「敬体ます」により、「にちがいありません」という表層形が定まる。HB3 は、後続要素に基づいて活用する語の活用形を自動決定・生成する機能を有しているため、一意に定まる活用形情報は Panzer の出力には含まれない。

3 文末変換

HB3 を用いて文末述語の表層文字列を復元できることはすでに述べたが、Panzer の出力の一部を書き換えてから HB3 で表層文字列生成を実行することにより、文末の書き換えが実現できる。

具体例を図３に示す。この図に示すように、「テ助動詞いる」を削除すると、「書いたにちがいありません」が生成できる。「テンス有標」を削除すると「書いているにちがいありません」が、「にちがいない」を削除すると「書いていました」が生成できる。最後の「敬体ます」を削除すると「書いていたにちがいない」が生成できる。削除だけでなく、置換も追加も可能である。たとえば、「敬体ます」を「敬体です」に置き換えると、「書いていたにちがいないです」が生成できる。「にちがいない」を「だろう」に置き換えると「書いていたでしょう」が生成でき、さらに末尾に「終助詞か」を追加すると、「書いていたでしょうか」が生成できる。

Panzer は、実はこのような変換を実現する必要性から生まれたシステムである。開発の直接の契機となったのは、話し言葉－書き言葉の自動変換、および、広告コピーの自動生成という、まったく異なる２つの応用タスクである。どちらも応用においても、以下に示すような文の文末形式を異なる形式に変換することが必要となった。

話し言葉－書き言葉の自動変換[3] では、ウィキペディアに記述されている人物エピソード文（書き言葉）を、対話エージェントが話しても違和感のない話し言葉に変換する。これを実現するために、体言止めの補完、伝聞モダリティの追加、丁寧体への変換などが必要となり、これらの変換を Panzer と HB3 を組み合わせて実現した。具体的には、次のような変換を行う。

趣味は足のネイルアート
→ 趣味は足のネイルアートだそうです
（ネイルアート＋だ＋伝聞そうだ＋丁寧）

この例では、体言止めである「ネイルアート」を内容語１語の文末述語と認定し、「判定詞だ」、伝聞を表す助

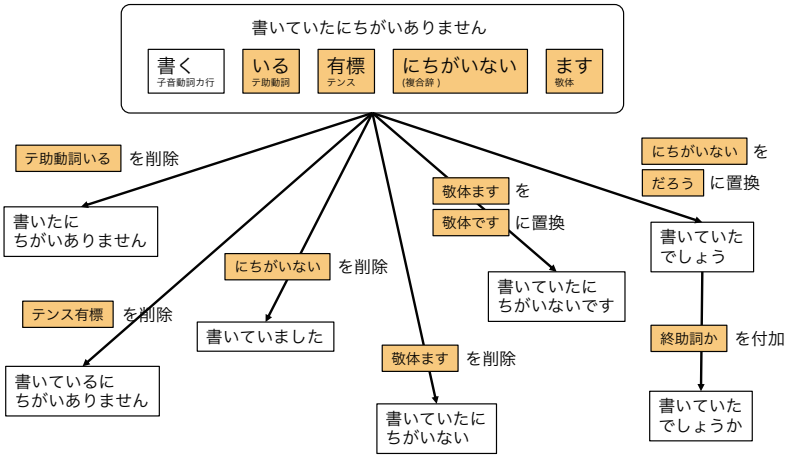

図３ 文末変換の例
Panzer の出力を書き換えると、色々な文末を生成できる

動詞「そうだ」、「丁寧」という３つのブリックを追加することにより、「ネイルアートだそうです」という形式に変換する。

広告コピーの自動生成[4] では、最小限の入力から、多くのバリエーションを生成するため、症状を記述する文から症状を問いかける文に変換することが必要となった。

腰が痛い

→ 腰が痛くはありませんか

（痛い＋は＋取立助動詞ある＋終助詞か＋丁寧）

日本語の文末述語には、テンス、アスペクト、モダリティ、肯定・否定・疑問、待遇表現などが集中する。そのため、文末述語の一部を書き換えるだけで、色々な変換が実現できる。

数年前から我々が取り組んでいることの一つに、小説の会話文の生成[5] がある。小説の会話では、それぞれの発話が誰の発話であるかを明示せずに、話し手に応じて口調を変え、誰が発話したかがわかるようにすることが多い。たとえば、

「フランス語を教えてください」

という依頼の発話を

「フランス語を教えろよ」

「フランス語を教えてくださらない」

のように文末述語の形式を変更すれば、読み手が受ける話し手の印象が大きく変化するため、登場人物の中の誰の発話であるかを暗黙的に伝えることができる。このような、話し手のパーソナリティーを感じさせる発話文の生成にも、文末の変換が活躍する。

4 おわりに

本稿では、Panzer による文末述語解析と、それを HB3 と組み合わせた文末変換について述べた。HB3 の設計思想の一つに、「できるたけ少ない記述で表層文字列の生成をプログラミングできるようにする」という方針がある。Panzer もその思想を受け継ぎ、文末述語の解析結果を、非常に簡潔な形式で出力する。この簡潔さが文末変換の実現を容易にしている。

参考文献

[1] 佐野正裕，佐藤理史，宮田玲．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．言語処理学会第 26 回年次大会，B6-3，pp1483-1486，2020.

[2] 佐藤理史．HaoriBricks3：日本語文を合成するためのドメイン特化言語．自然言語処理，Vol．27，No.2，pp411-444，2020.

[3] 柳将吾，佐藤理史．ウィキペディアから抽出した人物エピソードの話し言葉への変換．言語処理学会第 26 回年次大会，C2-3，pp437-440，2020.

[4] 平良 裕汰朗，佐藤 理史，宮田 玲，今頭 伸嘉．ダイレクト広告コピー文の分析と自動生成．言語処理学会 第 25 回年次大会 発表論文集，pp406-409，2019.

[5] 木村 遼，夏目 和子，佐藤 理史，松崎 拓也．発話表現文型辞書を利用した多様な発話文生成機構．2018 年度人工知能学会全国大会（第 32 回），2E2-02，2018.