特許文における入れ子構造の調査

Investigation of Nested Structure in the Patent Sentences

山形大学工学部技術部情報技術室技術員 髙橋　尚矢

PROFILE: 1989 年生。2012 年山形大学工学部卒。同年山形大学工学部技術部情報技術室に採用。

naoya_takahashi@yz.yamagata-u.ac.jp　TEL 0238-26-3339

山形大学大学院理工学研究科教授 横山　晶一

PROFILE: 1949 年生。1972 年東京大学工学部卒。同年電子技術総合研究所入所。1991 年同所知能情報部自然言語研究室長。1993 年 4 月より山形大学。現在大学院理工学研究科（情報科学分野）教授。工学博士。アジア太平洋機械翻訳協会（AAMT）Japio 特許翻訳研究会副委員長。

yokoyama@yz.yamagata-u.ac.jp　TEL 0238-26-3336

1 はじめに

近年、国際的な特許の共有化に伴い国際特許の申請数も増加し続けている。特許文の検索や翻訳などの作業には多くの人手が必要であり、そのため作業を自動化または半自動化することが求められている。これらを解決するためには特許文中に含まれる情報を的確に抽出することが要求される。したがって、特許文に対する正確な係り受け解析が不可欠である。

特許文の課題や解決手段の部分は、200 文字を超える長大な一文になることが多い。しかも単語同士の係り受けが複雑であり、意味が明確でないことがある。これは特許文独特の記述や専門用語の多さなどが原因である。そのため通常文に比べると係り受けが曖昧になりやすく、解析の誤りが発生しやすいという特徴がある。

本稿では、係り受けの中でも入れ子構造（nested structure）を取るものに着目した。係り受けの複雑化の要因の一つである入れ子構造を正しく解析できれば、機械翻訳する際に長文を一気に翻訳するのではなく小さなまとまりごとに翻訳でき、精度が高くなると考えた。

今回は“主辞（head word）＋英数字”という構成の語に着目し入れ子構造の検出を目指す。理由として並列構造をつくる要素には“装置 1”のような先程の構成を持つ語が頻出するからである。またこのような語は係り受け先の特定が難しいため長文で機械翻訳しづらい。よって小構造ごとに機械翻訳するという前提のもとで入れ子構造の検出をおこなう方針である。

2 関連研究

松山ら [1] は特許文ではなく法令文書を対象にした並列構造解析をおこなっている。あらかじめ法令文書の重要語を決め、それらをもとに並列構造を検出した。具体的には並列構造を pf1（前方句）、key（並列となるキーワード）、pb（後方句）と定義し、この型に当てはまる部分を検出するというものである。しかし階層的並列構造は検出が難しい、前方句と後方句の長さが違いすぎると並列構造の検出に失敗しがちである、など改良の余地がある。また法令文書では並列を意味する語が細かい規則で設定されている [2] のに対し、特許文ではそのような規則が無いため並列階層の特定が困難である。

我々は、昨年は特許文の並列構造解析について主辞（昨年度は“接尾辞”と表記したが主辞が正しい）に着目した手法 [3] で解析を試みた。特許特有の主辞を抜き出しそれをもとに係り受け修正システムを作成し一定の結果が得られた。ただし主辞を限定したことにより全体の誤

り修正の精度はさほど上がらなかった。新たに発見した問題点としては入れ子構造が正しく解析されなかった点があげられる。今回はその入れ子構造問題に取り組む。

【解決手段】縦壁からなる取付基部１に垂直回転自在に連結され、前傾姿勢で床面３上に収容可能な手摺体２と、一端非拘束状態で手摺体２に折り畳み可能に連結され、床面３上での起立姿勢において手摺体２をほぼ水平な使用姿勢に保持可能な支持脚４と、支持脚４の折り畳み動作を規制し、該支持脚４の起立姿勢を維持する拘束部５とを有して構成する。また、収納姿勢において、支持脚４の非拘束端４ａが手摺体２先端より後方に位置するように構成する。

図１　実験で使用した特許文

3 入れ子構造

図１に示したのは公開番号特開 2003 － 176610 の文章から【解決手段】部分を一部抜粋したものである。今回の実験では下線部に注目した。このような入れ子構造を持つ特許文をKNP[4]で解析したものを図２に示す。この解析結果では独自性の強い複合語をうまく扱えていないほか、中央の四角で囲った箇所（「支持脚４と、支持脚４の」の部分）が並列構造を間違って捉えているという問題がある。本来であれば「拘束部５“とを”有して」とあるように２つ以上の事柄を並列させていなければならないが、この解析結果では拘束部５のみをピックアップする形である。支持脚４と拘束部５の並列関係が成り立たない出力となっているため不適切な出力となる。

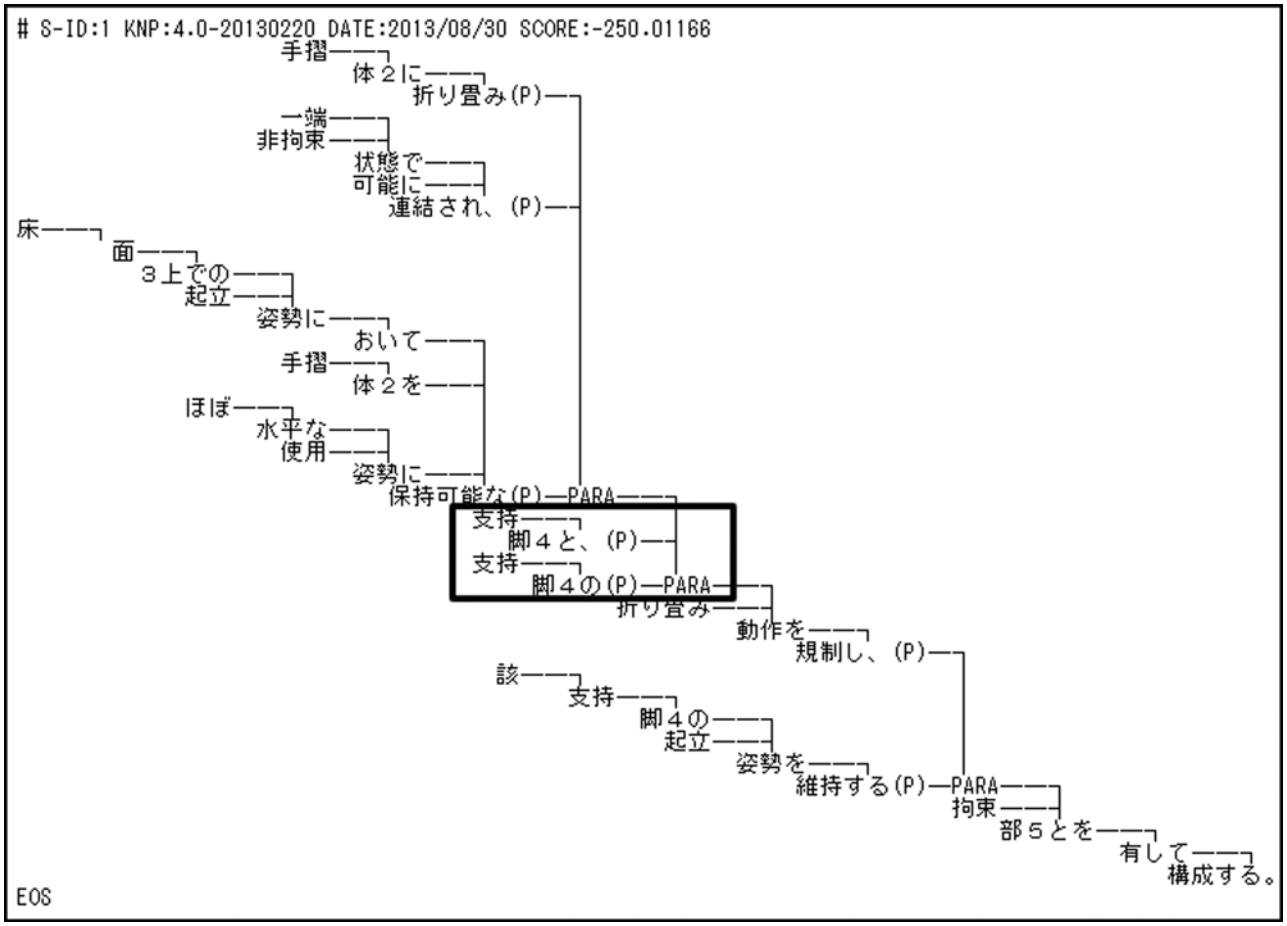

図２　KNP による特許文の構文解析結果

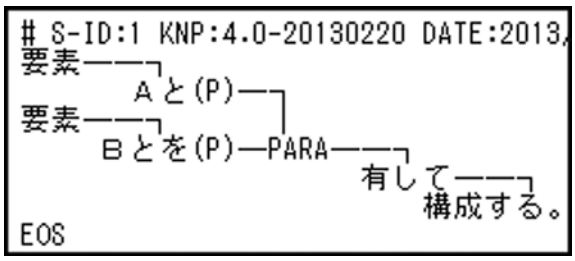

図３　単純化した文の構文解析結果

もしこれが簡潔な文ならば図３のように正しく解析できることは確認済みである。図３の実験では「一端非拘束状態で手摺体２に折り畳み可能に連結され、床面３上での起立姿勢において手摺体２をほぼ水平な使用姿勢に保持可能な支持脚４」を『要素 A』に、「支持脚４の折り畳み動作を規制し、該支持脚４の起立姿勢を維持する拘束部５」を『要素 B』に置き換えて構文解析した結果である。

上記の結果を踏まえ、方向性としては長い特許文を短く単純にする手法を探る。詳しい調査はまだおこなっていないが、“主辞＋英数字”の構造を取る語はそれ自体が並列構造に絡みやすいように見える。並列の接続詞や助詞を伴う“主辞＋英数字”がたくさん現れる箇所は入れ子構造も出現しやすいと考えられる。その箇所に小さなまとまりを発見できれば入れ子構造になっていると推測できる。

並列構造であるかの予測には主辞に付随する英数字を判断材料にする。たとえば英数字が同じもの同士は直接の並列関係にはなりにくい。このような規則ベースを作成しコスト削減も目指す。現在検討中の規則は“同じ主辞＋異なる英数字”の場合である。この構造の語間に並列の接続詞や助詞がある場合は並列構造である可能性が高い。一方で並列品詞が無い場合は修飾・非修飾の関係ではないかと推測できる。意味的な評価が入ってくるため機械学習も取り入れたい。

4　今後の方針

今回の調査で、構造解析システムは長文の解析が不完全であることが改めてわかった。特に並列構造を正しく捉えることがシステムでは難しい。複合語など未定義語の問題は辞書の拡充などで対応できるかもしれないが、並列構造の場合においては長文だからこそ起こりうる失敗例だといえる。ただし最小の並列構造ごとに分解すれば構造解析システムでも正しく解析をおこなえる可能性を示せた。

今後は特許文における入れ子の並列構造を自動で判断、抽出し小構造で解析する手法を検討する。小構造で解析したものどうしを組み合わせていき、最終的にはひとつの特許文解析を復元する。手法としては入れ子構造の判定をある程度人手で付与したのち教師あり学習で自動化することを検討している。

参考文献

[1] 松山宏樹、白井清昭、島津明：法令文書を対象にした並列構造解析、言語処理学会 第 18 回年次大会 (2012)

[2] 岩本秀明、長野馨、永井秀利、中村貞吾、野村浩郷：法律文における並列構造の特徴とそれに基づく制限言語モデルについて、情報処理学会自然言語処理研究会 (1993)

[3] 横山晶一：接尾辞に着目した特許文の並列構造解析、Japio Yearbook(2012) pp.250-253

[4] 日本語構文・格解析システム KNP：（http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP）