

接尾辞に着目した特許文の 並列構造解析

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

☎ 0238-26-3336

1 はじめに

特許文の課題や解決手段の部分が複雑であり、200文字を超える長大な文になるとともに、その係り受け構造も複雑になるということは、すでに何度も言及してきた [1-4]。

これまで、特許文解析に特徴的な、複雑な係り受け構造を解明するため、並列接続詞 [5] や、並立助詞 [6] について調査し、誤り自動修正システムを構築してきた。

本稿では、並列に重要な役割を果たす名詞を、広く接尾辞としてとらえることによって、特許文の係り受けを修正するシステムについて述べる。ここでの記述は、主として [7] に基づき、その内容に加筆したものである。

2 特許文における並列要素

特許文の請求範囲、特に解決手段は、全体が一つの文で書かれることが多く、長大で複雑な文になりやすい。多くは、長い修飾句を有する並列構造になっている。これまで、この修飾句の構造を解明する手がかりとして、並列接続詞、並立助詞について調査し、係り受けの誤りを修正するシステムを構築してきた。本稿では接尾辞に着目するが、それぞれの特徴について簡単に述べる。

(a) 並列接続詞

並列接続詞には、「または」、「もしくは」、「および」、「ならびに」などがある。法律用語では、これらの間に階層

関係を設けて、法律の条文の曖昧性をなくす方策を取っている [8] が、我々が特許文を調査した結果では、法律の条文のような階層性は特に見られなかった [5]。

(b) 並立助詞

並列を表す助詞として、「と」、「や」、「か」などがある。特許文では、「AとBとの～」といった形でよく使用される。これらの構造をとらえて、係り受けの誤り修正に使用したことがある [6] が、より複雑な構造に対しては、余り効果を発揮できなかった。

(c) 接尾辞

接尾辞とは、一般に、語の末尾に付けて、意味を加えたり、品詞を変化させたり、丁寧さや数など、文法上の変化をもたらしたりする形態素のことを言う [9]。たとえば、「A部と、B部と…」における「部」や、「C層およびD層では…」における「層」が相当する。図1に、特許文の解決手段に書かれた例を示す。この図で、太字で下線を引いた「部」が並列句になっているが、その他の下線の「部」は、この並列構造に寄与していない。この構造は、[]でくくった「と」や「とを」という並立助詞からとらえることができる。

しかしながら、一般的な意味での接尾辞（形態素解析システムでは、「名詞、接尾」といった表示で示される）では、後述するように、特許文の並列構造をすべてとらえることはできない。本稿では、特許文特有の並列構造に係る名詞の概念を拡張して、並列構造に係る名詞を接尾辞としてとらえることによって、並列構造をより明確にすることができることを示す。

[解決手段] 電力線通信装置は、電力線に接続された他装置との間で装置情報を送受信する初期接続部[と]、前記初期接続部が送受信した装置に基づいて、自装置が通信可能相手情報を作成する通信可能相手情報作成部[と]、前記通信可能相手情報作成部が作成した自装置が通信可能相手情報を他装置との間で送受信する通信可能相手情報送受信部[と]、前記通信可能相手情報送受信部が送受信した自装置および他装置の通信可能相手情報に基づいて、自装置が通信可能な他装置を階層状にマッピングした登録情報を作成する登録情報作成部[と]、前記登録情報作成部によって作成された登録情報を他装置との間で送受信する登録情報送受信部[とを]更に設けた。

図1 特許文における並列構造に係る接尾辞の例 (公開番号：特開 2010-21954)

他方部に土盛装置と畔形成装置と
 *0 2D 1/2 0.644114
 他方 名詞, 一般, ***, 他方, タホウ, タホー
 部 名詞, 接尾, 一般, ***, 部, ブ, ブ
 に 助詞, 格助詞, 一般, ***, に, ニ, ニ
 *1 2D 2/3 0.000000
 土 名詞, 一般, ***, 土, ド, ド
 盛 名詞, 固有名詞, 地域, 一般, **, 盛, サカリ, サカリ
 装置 名詞, サ変接続, ***, 装置, ソウチ, ソーチ
 と 助詞, 並立助詞, ***, と, ト, ト
 *2 -1D 2/3 0.000000
 畦 名詞, 一般, ***, 畦, アゼ, アゼ
 形成 名詞, サ変接続, ***, 形成, ケイセイ, ケイセイ
 装置 名詞, サ変接続, ***, 装置, ソウチ, ソーチ
 と 助詞, 並立助詞, ***, と, ト, ト
 EOS

図2 南瓜の解析例

3 研究方法

3.1 資料

解析用の資料として、AAMT/Japio 特許翻訳研究会 特許情報データベース [10] を用いる。このデータベースは、2004 年分の公開特許公報をすべて収録したもので、特許件数約 34 万件、文章数では約 101 万文ある。

3.2 接尾辞の抽出と分類

上記資料の文を南瓜 [11] に入力して、係り受け解析結果を出力する。図2に、南瓜による解析例（一部省略）を示す。図2で、数字の並んだ行は、文節とその係り先である。「0 2D」は、「他方部に」が、「畔形成装置と」に係っていることを示す。また、各形態素の右側の記述は、それぞれがどんな品詞となっており、読み方等がどのようなになっているかを示している。

図2で、下線を引いた「部」は、既述のように、接尾辞であるが、第2、3文節に現れる「装置」も、並列的な接尾辞と見なすことができる。このような名詞を、本稿では、「特許特有接尾辞」と呼ぶ。

3.3 特許特有接尾辞の抽出

本稿では、同じ並列修飾句に並列キーをはさんで並列要素として複数出現する頻度の高い名詞（一般名詞）を「特許特有接尾辞」とする。並列のキー [12, 13] としては、「と」、「や」、「か」、「、」、「・」、「かつ」、「および」、「または」、「ならびに」、「あるいは」、「もしくは」、「だけで（は）なく」を用い、数値とアルファベットを除外して、上記資料の約 101 万文を調査した。

並列句に現れる高頻度名詞(339716特許, 1013582文)

手段	名詞, 一般	23523 回
装置	名詞, サ変接続	19906 回
工程	名詞, 一般	12305 回
方法	名詞, 一般	10683 回
情報	名詞, 一般	7229 回
基	名詞, 一般	6579 回
データ	名詞, 一般	4671 回
部材	名詞, 一般	4534 回
ステップ	名詞, 一般	3967 回
位置	名詞, サ変接続	3674 回

図3 特許特有接尾辞の一部

図3に、抽出した特許特有接尾辞の一部を示す。本研究では、上位100語、出現頻度413回以上の語を用いる。

3.4 分野別の出現頻度

特許には、国際的に用いられている文献の技術内容によるカテゴリがあり、記号によって分類されている。「セクション」、「サブセクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の順に分類が細かくなっている。

セクションは、A（生活必需品）、B（処理操作、運輸）、C（化学、冶金）、D（繊維、紙）、E（固定構造物）、F（機械工学、照明、加熱、武器、爆破）、G（物理学）、H（電気）の8分野に分かれている。

図4に、セクションCの特許データについて調査した結果を示す。すべての特許文を対象にした場合に高頻度にならなかった「原子」、「酸」などの名詞が上位に来ていることが分かる。

化学、冶金分野の高頻度名詞（27969特許、76517文）		
基	名詞，一般	3729回
以下	名詞，非自立，副詞可能	1941回
工程	名詞，一般	1844回
樹脂	名詞，一般	1654回
方法	名詞，一般	1613回
原子	名詞，一般	633回
酸	名詞，一般	580回

図4 C（化学、冶金）分野における特許特有接尾辞

4 係り受け修正システム

特許特有接尾辞を用いたシステムを作成すると、従来はできなかった、並列構造の中に入れ子構造的にまた並列構造を含む文の係り受け誤りを修正することができる。

図5に、このようなシステムの出力例を示す。図の左側の番号は、文節番号を示し、その右側の2D、3Dという番号は、係り先の文節を示す。係り受けの修正を行っ

0 28D		また、
1 2D		その
2 3D		ための
3 28D		制御冷却装置は、
4 5D		圧延直後の
5 6D		鋼板の
6 7D		面温度分布を
7 8D		測定する
8 15D	(8 17D)	温度測定〔装置〕と、
9 11D		冷却水ヘッダーと
10 11D		これに
11 12D		接続された
12 13D		ラミナー状の
13 14D		冷却水を
14 15D		供給する
15 16D		ノズルとを
16 17D		含む
17 23D	(17 27D)	冷却<〔装置〕>と、
18 19D		所定の
19 20D		計算プログラムに
20 23D		したがって
21 22D		鋼板の
22 23D		面温度分布を
23 25D		均一化するように
24 25D		冷却水量を
25 26D		制御する
26 27D		冷却水量の
27 28D		制御<装置>とを
28 -1D		備える。

図5 システムによる係り受け誤りの修正成功例

た場合には、(8 17D) のようにカッコに入れて表示し、係り元の接尾辞を〔〕で、係り先の接尾辞を<>で囲んだ。

この図の文は、文節8と文節17の並列構造の中に、文節9と文節17の並列構造を入れ子的に含んでいる。このような文でも、特許特有接尾辞（ここでは「装置」）

に着目することにより、係り受けを修正することができる。

5 評価と考察

2004年分のデータから、接尾辞を含む並列構造文500文を無作為に選び、システムにかけた結果、どの程度誤りを修正することができたかを評価した。その結果を表1に示す。

表1 システムによる係り受け誤り修正結果

	正→正	正→誤	誤→正	誤→誤	合計
件数	318	16	97	69	500
%	63.6	3.2	19.4	13.8	100

この表で、「正→正」は、もともと並列構造を正しく解析できていたものをこのシステムにかけてもやはり正しく解析できたことを示し、「誤→正」は、解析誤りを修正できたことを示す。

表1からわかるように、誤り166件(97+69件)のうち、97件(58.4%)を修正できた。これは過去の結果[4, 5]である55.7%よりもわずかではあるが改善されている。

6 まとめと今後の課題

本研究では、接尾辞を考慮することによって、並列構造の解析誤りを修正する試みについて述べた。特許特有接尾辞の数を限定したことが、誤り修正の程度が期待ほど上がらなかった原因と考えられる。今後は、まず、語数の増加をはかる。次に、シソーラスを組み合わせたり、接尾辞の直後に来る数値や記号に着目したりすることによって、さらに精度の向上を目指す予定である。

謝辞

特許データベースをご提供いただいた Japio に感謝致します。

参考文献

- [1] 横山晶一、高野雄一：語のグループ化を用いた特許文動詞の自動訳し分けに関する調査、Japio Yearbook (2011) pp.234-237
- [2] 横山晶一、高野雄一：特許文の英語への訳し分けと述語の関係、Japio Yearbook (2010) pp.274-279
- [3] 横山晶一：特許文の英語への訳し分けと格フレームとの関係、Japio Yearbook (2009) pp.262-265
- [4] 横山晶一：動的シソーラスを用いた特許文の解析システム、科学技術研究費成果報告書(2007～2009)
- [5] 横山晶一：特許文における接続詞と係り受けの構造、Japio Yearbook (2008) pp.68-73
- [6] 横山晶一：特許文解析誤り自動修正システムと正確な翻訳のための特許文の分割、Japio Yearbook (2007) pp.228-233
- [7] 坂本和磨：接尾辞に着目した特許文の並列構造解析、山形大学工学部情報科学科卒業論文(2012)
- [8] 田島信威：最新法令用語の基礎知識(三訂版)、ぎょうせい(2006)
- [9] 新村出編：広辞苑(第六版)、岩波書店(2008)
- [10] AAMT/Japio 特許翻訳研究会特許データベース(2004)
- [11] 奈良先端科学技術大学院大学：係り受け解析器「南瓜」
- [12] 山村広臣、菅沼明、牛島和夫：日本語文における並列構造の簡便な推定法および推敲支援への適用、情報処理学会第52回全国大会(1997)
- [13] 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)