並列接続詞を含む特許文の係り受け修正システム*

横山　晶一†　　小野　裕太‡　　橋本　力†
（†山形大学大学院理工学研究科）（‡山形大学工学部）
yokoyama@yz.yamagata-u.ac.jp

１．はじめに

特許文の請求範囲や詳細は、習慣的に 200 文字を超えるような長い一文で書かれ、人間が見ても係り受け関係が複雑で、分かりにくい文になることが多い。

特許文は、申請時に、過去の類似特許を検索、調査するなど、類似情報を的確に把握する必要がある。また、機械翻訳の際にも、係り受け関係が正しく捉えられていることが必要になる。

すでに、特許文の係り受け解析を行った時の誤りを分析して、その結果を自動修正するシステムについて発表した[1,2]。これは、並列構造に、助詞「と」を含む文についてのものであった。

本稿では、「または」、「もしくは」といった並列接続詞を含む特許文を分析し、その係り受け関係を修正するシステムについて述べる[3]。並列接続詞は、法令では、その使用法や上下関係が定まっている[4]が、実際の特許文を分析した結果では、必ずしもこの規則に従っていないことが判明した。

ここでは、その分析結果について述べるとともに、係り受け解析システムで誤った係り受け関係が生じた場合に、それを自動修正するシステムを具体例とともに示す。これは、従来のシステム[1, 2]を基本的には拡張、改良したものであるが、特に並列接続詞を含む並列構造の修正を意図したものである。

システムでは、分析した文に対して、日本語語彙大系[5]の意味階層を用いたり、接尾辞や数詞を用いて自動修正を行う構造になっている。分析した文の約３分の２程度を修正することができた。また、このシステムを用いて、並列接続詞「もしくは」や「または」を含む別の文を試行したところ、誤った文をある程度修正できることが判明した。

* Correction system for dependency of patent sentences including parallel conjunction, YOKOYAMA Shoichi, ONO Yuta, and HASHIMOTO Chikara, Yamagata University

２．特許文の特徴と並列構造

(11)【公開番号】特開 2001－219655
(43)【公開日】平成 13 年８月 14 日
(54)【発明の名称】熱転写記録媒体及び画像形成方法
(21)【出願番号】特願 2000－30516
(22)【出願日】平成１２年２月８日
(72)【発明者】【氏名】椎名 義明
(57)【要約】【課題】インキの滲みによる解像力の低下を防ぐ事、また、ワックスを使用していると転写した画像を手で擦ったような場合の耐久性が足りず、画像が取れて無くなり易いことなどの点を改善して耐久性を増す事、そして特に、感熱転写の際の感熱転写シート基材の剥離時における熱転写記録層の箔切れ性（膜切れ性）が良く、且つ転写画像の光学濃度も高い事、これらを同時に充分達成することのできる熱転写記録媒体（感熱転写リボン）を提供する。【解決手段】支持体２上に少なくとも着色顔料と有機樹脂バインダーと無色又は淡色の微粒子とを主成分とする組成物から形成された熱転写記録層３が設けられた熱転写記録媒体１において該熱転写記録層が膜厚０．５～１．０μｍの範囲にあり、前記有機樹脂バインダーが平均分子量１００００～２００００の範囲の塩化ビニル－酢酸ビニル共重合樹脂である。

図１　特許文の例[6]

特許文の特徴は、前述のように、並列構造を多用した長い文である。図１に、例を示す。この図の「要約」の「課題」や「解決手段」は、特許文としては比較的分かりやすい方であるが、長い並列文が多く含まれている。本文の請求範囲や詳細では、もっと長く、分かりにくい文が頻出する。日本語の一般的な文では、だいたい 20～100 文字程度が一文であるが、特許文では 200 文字を超える。

このような長い文の並列構造の解析には、文

節同士の類似性を発見することによってうまく構文解析する[7]手法が開発されてきた。これは、通常の長い文に対しては有効であるが、特許文のように、長い名詞句でつながった並列構造については、必ずしも有効ではない場合がある。

また、文書の定型的表現を手がかりとして、特許請求項を構造解析した研究もある[8]。これは、特許の文字列や、「であって」のような手がかり句を用いて、特許請求項の構造を解析するもので、これらの構造を持った特許文解析に有効であることが示されている。

我々のグループは、特許文の中に、長い名詞句の並列構造が多いことに着目し、「A と B との C」（A, B, C は名詞）といったパターンを持つ特許文における係り受け解析（「南瓜」[9]を使用した）の誤りを修正するシステムを構築した[1,2]。図２に一例を示す。

0　1D 製造設備、
1　2D 検査設備の
2　3D 各装置個別の
3　4D <<3 7D>> データ収集と
4　7D データ解析を
5　6D 下位の
6　7D ネットワーク上で
7　8D 可能とし、

図２　名詞の並列構造の修正例[1]

図２では、一番左に文節番号を示している。その後の 1D, 2D といった番号は、係り先の文節を示す。図では、並立助詞「と」を検出した時点で係り先を 4 から 7 へ修正したことを示している。

その他に、読点や副詞節の情報をもとに、長い特許文を、正しい係り受けが可能な範囲に分割することも試みた[10]が、並列構造を効率的に分割するには至っていない。

３．法令用語における並列接続詞

法律の条文では、解釈の曖昧性を避けるために、並列接続詞に上下関係を設けて、その使い方を定めている[4]。

たとえば、”or”に相当する接続詞「または」と「もしくは」（表記は「又は」、「若しくは」と書かれる場合もある）では、「または」の方が優先される。

(例 1)［（A もしくは B）または C］
(例 2)［（A または（B もしくは C）］

つまり、「または」と「もしくは」が両方出現した場合、「もしくは」の方を先にまとめ、「または」を上位でまとめるという規則が定まっている。

我々はこの点に着目し、特許文の構造にも類似の点があるのではないかという観点から、予備的な調査を行った。

４．特許文の並列構造の調査

予備調査として、特許データベース[11]の中から、「または」、「もしくは」を両方含む文のデータを集計した。このようなパターンは、特許データの中では意外に少なく、3900 件の特許を調査した中で 43 件しかなかった。

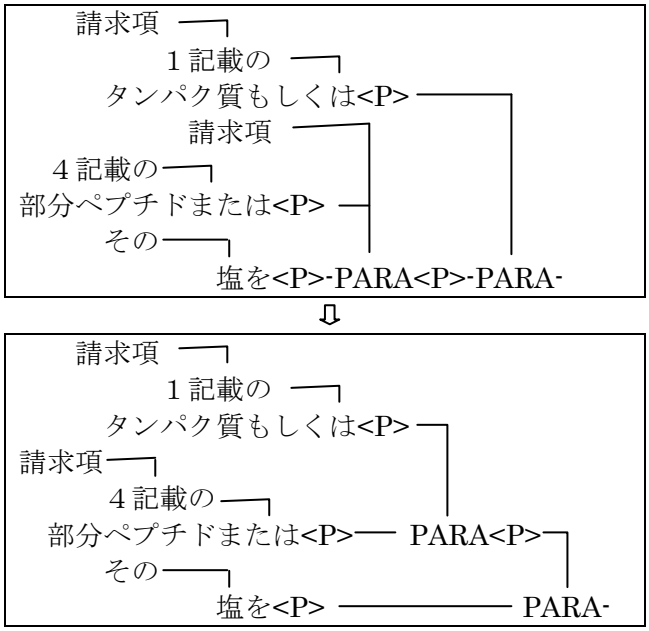

図３　法令用語の規定に従う並列構造の例

図３は、法令用語に従った並列構造を持つ特許文の一部を示したものである。図の上の部分は、KNP [12] の出力を示す。図の下の部分は、正しいと思われる構造に人手で修正したものである。本稿では、予備調査の結果、形態素解析には、並列構造により強いと考えられる KNP を係り受け解析に用いている。

この図で分かるように、「A もしくは B」でまとまった句を作り、それが「または」と並列構造を成している。

しかしながら、逆のケース（例 2 に示した形式）の場合には、必ずしも法令規則には当てはまらないことが判明した。

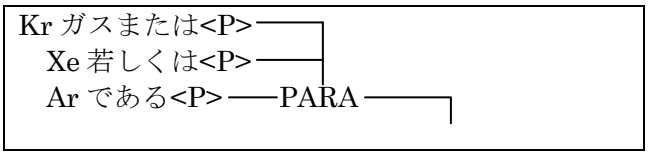

図 4　法令用語の規定に当てはまらない例

図 4 は、そのような例である。もし法令用語の規定に従っていれば、「Xe 若しくは」と「Ar」がまとまらなければならないが、多くの特許文では、この例のように、「または」、「もしくは」を同等の接続詞と見なして、最初の語から係る形を取っている。

法令文では、上記の優先関係が守られていることは、調査の結果分かっている[13]が、特許文では、奨励はされている[14]ものの、多くの場合守られていないことが、この予備調査の結果判明した。

５．「若しくは」における係り受け構造誤りの分類

「若しくは」を含む並列構造を、KNP で解析した結果、係り受けが誤っていると考えられる文を 153 件抽出し、誤りの性質ごとに分類した。表 1 に分類結果を示す。その他に分類不明のものがあったが、それは表 1 からは省いてある。

表 1 「若しくは」の並列構造誤りの分類

分類	(a)	(b)	(c)	(d)	(e)
誤り数	113	19	12	4	5
割合(%)	73.9	12.4	7.8	2.6	3.3

分類を以下に示す。

(a) 「A 若しくは B」の形で、A と B が近い意味階層で対比している場合

調査した中では、この形が最も多く、全体の半数を超える。具体的には次のような前後に長い修飾部を伴う語句の対比になっている。

(例 3) 生ゴミが収容される有底筒状のゴミ容器と、～と、～と、…モータを連続的もしくは断続的に駆動させて～

(b) 「A 若しくは A の B」の形になっている場合

(例 4) 培養液◆▶培養液から得られた菌体

(c) 「A 若しくは B 又は C」の形で、A,B,C が並列になっている場合

(例 5) 大きさ、若しくは色又は模様の変化

(d) 「A 若しくは B」で、A と B が離れていて、同じ文節のつながりから判断できる場合

(例 6) 熱電素子に供給する電圧値を最低電圧とするか、もしくは、前記熱電素子に供給する電圧値を上昇させる…

この場合には、下線部が並列であると判断できる。

(e) 「～の A、若しくは～の A」の形の場合

(例 7) 前記連結シール部の有効長さ、若しくは前記円筒シール部の有効長さ

６．係り受け誤り修正システム

前節の分析結果に基づいて、係り受けの誤りを自動修正するシステムを構築した。このシステムでは、「若しくは」を検出すると、その直前にある品詞を判別して、それに応じた処理を行う。

(1) 直前の品詞が名詞の場合(前節(a)に対応)

ここでは、まず日本語語彙大系[5]の意味階層の情報を得る。後ろの文の名詞と比較して、同じ意味階層が得られたならば、誤りを修正する。下から 3 階層以内で同じ意味番号が現れたならば、それも修正する。

(例 8) 和［1 名詞/1000 抽象/2422 抽象的関係/2443 関連/2476 均衡・不均衡/2477 均衡］
差［1 名詞/1000 抽象/2422 抽象的関係/2443 関連/2458 異同/2461 均衡］

この例では、「関連」の部分が一致するので並列と判定している。

(2) 接尾辞またはそれに類する名詞の場合

後ろの部分にも同じ接尾辞があれば、さらに先を見て、並列性を判定し、修正を行う。名詞の場合でも、

(条件 1) A もしくは A の～　(前節(b)に対応)
(条件 2) A もしくは A から～

という形になっていれば、その次を見て修正の対象とする。

(例 9) 培養液、若しくは培養液から得られた菌体
(例 10) 多結晶若しくは単結晶の…

例 9 は、上の条件 2 に合致するので、さらに先を見て、「培養液」と「菌体」が並列であると修正するが、例 10 は、条件は満たさず、通常の並列性が現れていると判断する。

(3) 数詞の場合

KNP では特に問題ないので修正を行わない。

(4) 形容詞の場合

後ろに形容詞が来れば、並列として修正を行う。

(5) 副詞の場合

日本語語彙大系にあり、意味階層が合っていれば修正する。そうでなければ形容詞と同じように副詞同士の並列性を判断する。

(6) 未定義語の場合

同じ語があれば修正する。

このような比較的単純なアルゴリズムでシステムを作り、表１の誤った文を入力して、修正できるかどうかを判定した。その結果を表２に示す。

表２　分類文に対するシステムの修正結果

	(a)	(b)	(c)	(d)	(e)	計
修正成功	87	5	9	2	0	103
修正不成功	26	14	3	2	5	50
修正割合(%)	77.0	26.3	75.0	50.0	0.0	67.3

全体では、約３分の２くらいの文が修正できているが、(c)以下の分類に対しては、まだアルゴリズムがきちんと働いていない。

このシステムを、無作為に選んだ「若しくは」を含む 744 文に対して適用した結果を表３に示す。これは、KNP の出力結果に、さらにシステムを適用して修正可能かどうか調べたものである。

表３　別の文に対するシステムの適用結果

	誤→正	正→正	正→誤	誤→誤
件数	93	513	64	74
割合(%)	12.5	69.0	8.6	9.9

誤りのうち 55.7%は修正できたが、一方で正しいものを誤って修正する場合も多い。

７．問題点と今後の課題

非常に初歩的なシステムしか構築できなかったので、問題点はまだ数多く残されている。並列接続詞は、前述のように、「または」、「もしくは」の他にも、「および」、「ならびに」などがある。これらについても、特許文では、法令文と異なり、前からの並列が多いことが確かめられている。

今後は、意味やオントロジーなどを使用して、さらに詳しい分析を行い、システムを拡張、改良していく予定である。

謝辞

特許データベースの提供や、特許関係の文献について種々ご示唆いただいた水谷直樹弁護士・弁理士、ならびに(財)日本特許情報機構(Japio)奥直也氏、大塩只明氏に感謝します。また、アジア太平洋機械翻訳協会(AAMT)と Japio による特許翻訳研究会のメンバーにも感謝します。
本研究は、科学技術研究費（基盤研究(C)課題番号 18500102）のもとで行われた。

参考文献

[1] 見年代茂大、横山晶一：特許文解析誤りの修正システム、情報処理学会第 69 回全国大会 6Q-3 (2007) pp.2-427-428
[2] YOKOYAMA Shoichi, KENNENDAI Shigehiro: Error Correcting System for Analysis of Japanese Patent Sentences, Machine Translation Summit XI, Workshop on Patent Translation (2007) pp.24-27
[3] 小野裕太：特許文の接続詞係り受け修正システム、山形大学工学部卒業論文 (2008)
[4] 田島信威：最新法令用語の基礎知識（３訂版）、ぎょうせい (2006)
[5] 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（編）：日本語語彙大系、岩波書店 (1997, 1999)
[6] 特許庁データベース http://www.ipdl.ncipi.go.jp/homepg_j.ipdl
[7] 黒橋禎夫、長尾真：並列構造の検出に基づく長い日本語文の構文解析、自然言語処理 Vol.1, No.1 (1994) pp.35-57
[8] 新森昭宏、奥村学、丸川雄三、岩山真：手がかり句を用いた特許請求項の構造解析、情報処理学会論文誌 Vol.45, No.3 (2004) pp.891-905
[9] 南瓜　奈良先端科学技術大学院大学松本研究室 http://chasen.org/~taku/software/cabocha/
[10] 吉田節行、横山晶一：特許文の機械翻訳における正しい係り受け判定のための文章分割、情報処理学会東北支部 2006 年度第 6 回研究会 (2007) B1-1
[11] Japio 特許データベース (2005)
[12] KNP　京都大学黒橋研究室 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
[13] 西村和夫：六法全書の統計、P は A か B の C か D である　http://www.komazawa-u.ac.jp/~kazov/Nis/study/law-andor.html
[14] 森智宏：ぱてんとさいと、条文用語解説http://patent.site.ne.jp/pa/terms.htm