特許文における接続詞と係り受けの構造

山形大学大学院理工学研究科教授
横山　晶一

PROFILE
1949年生。1972年東大工学部卒。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教授(情報科学分野)。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

1 はじめに

特許文の請求範囲や詳細が、日本語文としては異例に長い約200字を超える文から成り、その係り受け構造が複雑であるということは、すでに何度か言及した[1,2]。我々のグループでは、特許文の係り受け構造を正しく把握するために、特許文中の大きな係り受け構造をとらえる方針で研究を続けている。

昨年度から今年度にかけては、法令文と特許文との対比を行い、特許文における並列接続詞の使用法を明らかにするとともに、解析の誤りを自動修正するシステムを作成した[3~5]。ここではその内容を簡単に述べる。以下の2～7節の内容は、ほぼ[3]に基づき、一部加筆したものである。

また、特許文を明確化するために、日本語を明晰に書くという動きもある。これは、言ってみればある種の制限日本語に相当する。法令文も厳密な規定が定められているという意味では、一種の制限日本語である。これらについても簡単に言及する。

2 特許文の特徴と並列構造

図1に、特許文の例を示す。特許文の特徴は、すでに述べたように、並列構造を多用した長い文である。この

(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]

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[10,11]。図2に一例を示す。

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。

```
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［10］

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[12]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[13]。

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階層的な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[14]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[15]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、「南瓜」ではなく、並列構造により強

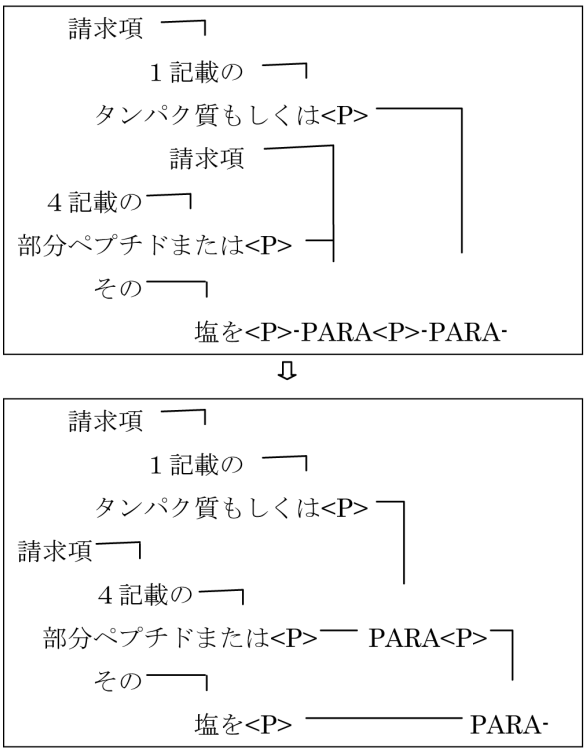

図3　法令用語の規定に従う並列構造の例

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。

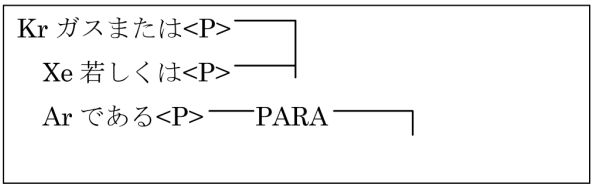

図4　法令用語の規定に当てはまらない例

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[16]が、特許文では、奨励はされている[17]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明らかになった。つまり、特許文では、日常言語と同様に、「または」や「もしくは」といった接続詞は、階層的な関係をなさずに、出現した順に、全く同等な語として用いられるということが判明した。

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。

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

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

(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[18]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。

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。

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。

表2 分類文に対するシステムの修正結果

有効文字数	10	25	50	100	200
相関係数	0.750	0.829	0.857	0.883	0.907
RMSE	2.308	1.918	1.777	1.617	1.428

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

7 係り受け修正システムの問題点

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、現在予備的な調査を行っているが、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。また、読点も、これらの並列接続詞と同様、同等な使われ方をしている（要するに、前からの並列に対して区別なく使われている）ことがほぼ明らかになってきており、これに対処したシステムを構築する必要がある。

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。

8 制限日本語への展望

もともとネイティブスピーカーでない人へ英語を教えるために、まず基本的な単語を設定して、その範囲内で読み書きができるようにするというbasic Englishなどの研究から、制限言語（controlled language）という考え方が出てきた[19]。機械翻訳との関連では、英語から多言語翻訳を行う場合に、前処理の過程を軽減する目的で、英語を制限するという研究がある[20]。日本語でも、長い文や語彙を制限したり、係り受け関係を制限したり、語句の関係を明確化するなどの試みが以前からも[21]、最近も[22]行われている。

法令文において、並列接続詞を階層化するのも一種の制限文法といえる。上記の試みは、たとえばマニュアルなどを書いて、人間に行わせてもなかなかうまくいかないので、システムで自動的に書き換え等を行う必要がある。これらは、書き換えや言い換えのシステムにもつながる重要な研究で、今後も種々の試みが行われると考えられる。

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] 横山晶一：特許文解析誤りの分類と自動修正の可能性、Japio Year Book（2006）pp.188-191

[2] 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 Year Book（2007）pp.228-233

[3] 横山晶一、小野裕太、橋本力：並列接続詞を含む特許文の係り受け修正システム、言語処理学会第14回年次大会発表論文集（2008）pp.87-90

[4] 小野裕太、横山晶一、橋本力：特許文の接続詞係り受け修正システム、2007年度情報処理学会東北支部研究会（2008）07-A-2-2

[5] 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）

[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] 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
[11] 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
[12] 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
[13] 田島信威：最新法令用語の基礎知識（3訂版）、ぎょうせい（2006）
[14] Japio特許データベース（2005）
[15] KNP　京都大学黒橋研究室　http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html
[16] 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
[17] 森智宏：ぱてんとさいと、条文用語解説 http://patent.site.ne.jp/pa/terms.htm
[18] 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）
[19] アジア太平洋機械翻訳協会編：機械翻訳―21世紀のビジョン（2000）pp.136-137
[20] Teruko Mitamura: Controlled Language for Multilingual Machine Translation, Proc. of MT Summit VII（1999）pp.46-52
[21] 長尾真、田中伸佳、辻井潤一：制限文法に基づく文章作成援助システム、情報処理学会自然言語処理研究会資料NL44-5（1984）
[22] 熊野明：特許文書処理と明晰日本語―機械翻訳の観点から―、産業日本語ワークショップ資料（2008）