

特許文解析誤りの修正システム

見年代茂大† 横山晶一‡

山形大学大学院理工学研究科

1. はじめに

特許文の請求範囲や詳細は、200 文字を超えるような長文が多く、その係り受けが複雑なことで知られている。日本語を母国語とする人間でも意味を読み取るのが困難である。特許文は申請時に同等特許が存在しないか膨大なデータベースの中から検索する必要がある。そのためには、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。また、国際的な特許の共有化に伴い機械翻訳の必要性も高まっている。したがって、これらの文に対する正確な係り受け解析が不可欠である。

本研究では特許文の係り受け誤りを抽出して、自動修正するシステムを開発し、その評価を行ったので報告する。

2. 特許文と係り受け誤り

日本の特許庁が平成 15 年（2003 年）に発行した公開特許公報全てのテキスト全文をまとめたデータ（DVD）[1]をサンプルとして使用し、同一特許の日本語文特許と英文特許の比較を行い、調査を行った[2,3]。その結果、日本語文における係り受け誤りが、英文への翻訳時にそのまま反映されることがわかった。このデータベースで使用された翻訳エンジンは不明だが、係り受け誤りが影響していることを上記の研究で明らかにした。

係り受け誤りの修正は、正しい日本語特許文の情報を取得することにつながり、係り受け誤りが原因でおきる文意の変化が修正されるなど、大きな意義をもつ。

3. 研究方法

3.1 係り受け誤り比較

サンプル内には、特許タイトルと要約部分を機械翻訳したデータが格納されている。この機械翻訳特許文と特許庁データベースにある人手翻訳による同一英文特許文[4]とを比較することで日本語の係り受け誤りの有無を判断した。

3.3 係り受け誤り調査・分類

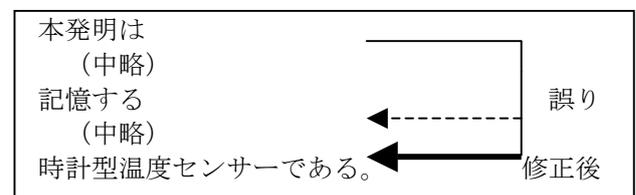
過去の研究での調査・分類を引用する[2,3]。サンプル[1]に格納されている項目は

“書誌的事項（番号・日付・出願人・発明者・発明の名称など）・要約・特許請求の範囲・発明の詳細な説明・図面の簡単な説明”などである。

この項目のうち要約と請求範囲を調査し、7つに分類した。代表的なものを以下に示す[5,6]。

(1) 特許特有表現

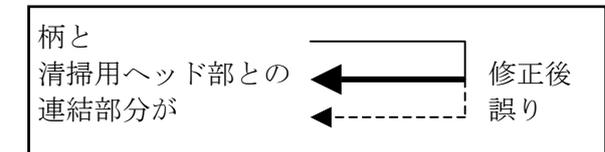
「本発明は～（中略）～Bである。」（AB 名詞）



上図では「本発明は」が「記憶する」に誤って係っている。正しくは「本発明は」が「時計型温度センサーである。」に係ることで正しい係り受けが成立する。

(2) 並列構造

「A と B との C が、」（ABC 名詞）



「A と」が「C が、」に係る誤り。他に「とを」「との」「とに」「と、」などがある。

正しくは「A と」が「B との」に係ることで正しい係り受けが成立する。

(3) 接続詞

(4) 名詞・動詞間の呼応

(5) 従属節間の係り受け

(6) 名詞修飾節

(7) 名詞+読点の並列構造

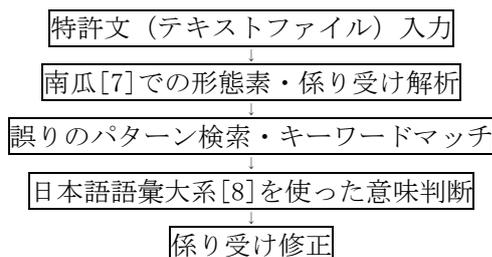
本研究では、分類上の性質と係り受け誤りの特性から(1)特許特有表現と(2)並列構造 について修正を行った。

4. プログラム

特許文を入力し、係り受け誤りがある場合は修正するプログラムを作成した。

基本的な流れは以下ようになる。

A System Correcting Modification Errors in Patent Sentences
†KENNENDAI Shigehiro Yamagata University
‡YOKOYAMA Shoichi Yamagata University



<入出力例>特許文 1 部分抜粋

「製造設備、検査設備の各装置個別のデータ収集とデータ解析を下位のネットワーク上で可能とし、」

- 0 1D 製造設備、
- 1 2D 検査設備の
- 2 4D 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、

この例文は「A と B とし、C」と分類した並列構造の誤りの修正である。本文の“と”と形態素解析結果で“データ収集と”の“と”が「助詞-並立助詞」と出力される。これに着目し、修正を行う。3 4Dは南瓜の結果表示で自分の文節とかかりうけ先の文節を表す。これを<3 7D>と修正を行うことにより対象の係り受けを修正した。

この例文では使用されていないが、パターンマッチに加え、日本語語彙大系を使用した意味判断で、パターンマッチで補えない修正を行った。

5. 評価とまとめ

表 1: サンプル分類

	総数 1228 件
特許特有表現	19
並列構造	209
接続詞	92
名詞+読点の並列構造	23
分類不能の誤り	85
正常 (係り受け誤り無し)	800

表 1 にサンプル[1]の 1 ファイル分にあたる 1228 件の特許文を上記のように人手で分類した結果を示す。そのうち特許特有表現と並列構造のうち「A と B との C が」という形のものをシステムで修正した結果を表 2 に示す。

表 2: 特許文サンプル集計数

	誤り	修正
特許特有表現	19	19
並列構造 (A と B との C が)	34	34

表 2 には示していないが、本研究の修正では「正しい係り受け」を誤って修正した例はない。

特許文特有表現の誤りは、全て修正することができた。並列構造「A と B との C が」についても分類された全ての誤りを修正することができた。

この修正では

「A と B との C」 (とに、との、と、を含む)

のようなパターンで、これは B の中にさらに「と並立助詞」を内包する文章、

「A と (A' と B') との C」等があった。この誤りは「A→C」が正しい係り受けであるが「A→A'」と係る誤りを修正することはできなかった。しかし「B' →C」の修正は行えた。

他の並列構造はたとえば「A、B を C する D と」がある。このパターンの問題点として「肉、卵、野菜、」のような“名詞+読点”の並列構造がある。日本語語彙大系や辞書などを用いて意味判断を行っても、検索の範囲・深さをどこまで行うかという問題があった。深すぎると正常な係り受けに対し誤った修正を行い、逆に浅い誤りは直せない。解決策として構文判断等の、複数の判断材料が必要であり、更なる研究が必要である。

「接続詞の係り受け誤り」は「並列関係の誤り」に次ぐ多さで「及び、又は等」の誤りがあったが、並列関係以上の精密な意味判断が必須であるので修正パターンを作成することが困難であり、本研究では見送った。

特許文は読点の位置や記述が記述者によってばらばらなため、係り受け誤りが存在するか否かも人手による主観判断を行う必要があった。主観による判断に加えて文意を読み取る必要があることが特許文研究を一層難解なものにしている。今後は特許文特有の構文解析を行い、特化した解析を行う必要がある。

参考文献

- [1]特許庁データベース、Japio(2005)
- [2]佐原洋輔：特許文の係り受け解析と修正、山形大学工学部卒業論文(2006)
- [3]横山晶一、見年代茂大、三戸部 矩倫：特許文の機械翻訳に与える日本語解析誤りの影響 平成 17 年度 AAMT/Japio 特許翻訳研究会報告書 (2006) pp. 45～52
- [4]特許庁データベース http://www.ipdl.ncipi.go.jp/homepg_j.ipdl
- [5]黒橋禎夫、長尾眞：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 No1(1994) pp. 35～57
- [6]南不二男：現代日本語文法の輪郭 大修館書店(1993)
- [7]南瓜 奈良先端科学技術大学大学院
- [8]池原悟他：日本語語彙大系、岩波書店(1997)