# 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

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。

## 1 はじめに

特許文が、他との差異を明確にして、新規性と進歩性を説明することが求められることは昨年度も述べた[1]。その中で、詳細部分はもちろんのこと、「要約」の「課題」や「解決手段」の部分などが、日本語の文には余り見られない長大な一文（約200字以上）になることが多く、その係り受けが、人間が見ても明確でないことについてもすでに言及した。

一方、特許文の検索や、翻訳などの作業に、コンピュータを用いて、自動的または半自動的に、人間の作業の肩代わりを行うことが求められている。これらがうまく行けば、関連する特許文の自動検索、キーワードの付与、英語などへの自動翻訳など、現在多くの人手を用いている作業をある程度軽減できることが期待される。

本稿では、これらの基礎として、まず、特許文に特徴的な長い文に対する係り受け構造を調査して、係り受け解析システムの誤りについて分類した結果について簡単に解説する[2,3]。その結果に基づいて、特許文の自動解析における係り受け誤りを解消するシステムを作成した[4,5]。この内容について述べる。

また、機械翻訳のためには、長い文を適切に分割できれば、係り受けや機械翻訳の改善に役立つことが知られている。その作業と、現在得られている知見についても述べる[6]。

本稿は、主として、昨年度のAAMT/Japio報告書に書いた報告[7]に基づき、その後の知見を加えたものである。

## 2 特許文の解析誤りの分類

Japioの要請で、アジア太平洋機械翻訳協会（AAMT）とJapioとの間で2003年9月に、AAMT/Japio特許翻訳研究会が作られ、特許翻訳の問題点について意見交換や研究の場が提供されることになったことも、すでに昨年の報告で述べてある。

本節で述べる研究成果は、その中から出てきた研究の一つであり、すでにいくつかのところで報告されている[2~9]。昨年度の報告と重複するが、ここでは、その中からいくつかの例を取って簡単に解説する。

手順は次のようになっている。

(1) 研究用に提供されている汎用の日本語係り受け解析ソフトである「南瓜」[10]を用いて、特許文の解析を行う

(2) 人間が解析結果を検証して、その中から係り受けの誤りを見つけ出す

(3) 誤りの種類を分類する

この結果、係り受けの誤りの種類がおおむね6種類（まだ分類しきれていないものを「その他」としてあるので、それを含めれば7種類）に分類された。ここではその中の代表的なものに限って、簡単に述べる。

(a)特許特有表現

**「本発明は～Aである」　（A：名詞）**

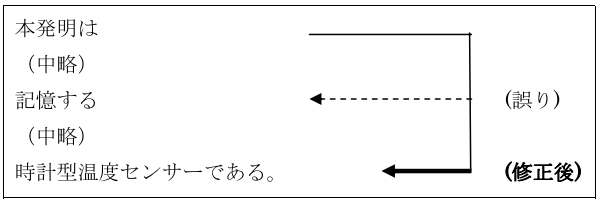


**図1　フレーズベース統計翻訳の動作例**

一般的に特許では、「本発明は」は最後の述語（動詞）に係る。しかしながら、南瓜の出力は、途中の動詞に誤って係ることが多い。これを上記のように修正する。

(b)並列構造

「AとBとのCが、…」　（A, B, C: 名詞）

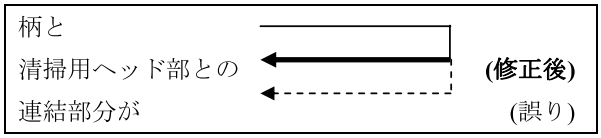


**図2　並列構造の誤り**

名詞の並列構造では、並列助詞の「と」の係り受けを誤ることが多い。上の図でも、「と」と「との」の並列が正しくとらえられていない。

その他には、接続詞、名詞・動詞間の呼応、従属節間の係り受け、名詞修飾節、名詞＋読点の並列構造などがある[3,4]。

## 3 解析誤りの自動修正システム

南瓜による解析結果の誤りを自動的に見つけて修正する簡単なシステムを作成した。このようなシステムには次のような利点がある。

(1) 正しい係り受け解析結果を得ることによって、関連特許の検索がしやすくなる

(2) 機械翻訳の誤りの原因の一つとして、これらの解析誤りが反映されるという問題がある。この問題を解消する一端になる

(3) 自然言語処理（日本語、英語など人間の用いる言語をコンピュータで処理すること）では、これまで正しい文入力を前提にしてきたが、誤りを修正するシステムによって、より発展的なシステムの可能性が開ける

システム構築に当たっては、まず、DVD[11]と、Web上の特許[12]とを比較対照した。前者は英文の部分が機械翻訳を使用しているが、後者は人間の手が入った翻訳である。この比較によって、前節に示したような分類が、機械翻訳の結果にある程度影響していることが確かめられた。

それに基づき、図3のような簡単なシステムを作成した。

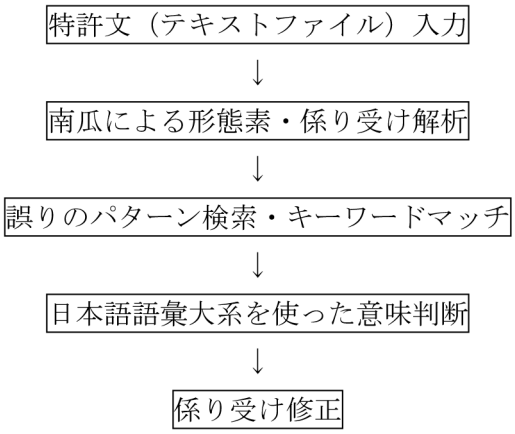


**図3　係り受け誤り修正システムの概要**

〈入出力例〉特許文一部分抜粋

「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[13]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。

**表1：サンプル分類**

| | | 1228件 | % |
|---|---|---|---|
| 特許特有表現 | | 19 | 1.5 |
| 並列構造 | 「AとBとのCが」 | 34 | 3 |
| | 「Aと、BをCするDと」 | 175 | 14 |
| 接続詞等、他の誤り | | 115 | 9.3 |
| 分類不能の誤り | | 85 | 7 |
| 正常（係り受け誤り無し） | | 800 | 65 |

表1にサンプル[11]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。

## 4 特許文の分割

すでに述べたように、特許文の解析結果の誤りが機械翻訳の結果にも影響している。そこで特許文を分割することにより係り受け解析誤りが機械翻訳に影響を与えることを回避できるか調査した[6]。以下に調査の手順を示す。

① 特許情報データベースから特許文を無作為に収集する
② 収集したデータに対し人手による分割を行う
③ それぞれの部分を機械翻訳に通す
④ WEB上の翻訳と比較することで評価する

ここでは、人間の手によって係り受け構造を考えて特許文を分割する。以下に人手分割と分割後の機械翻訳の一例を示す。

例：2003−000271

【課題】 アポトーシス制御や糖代謝制御シグナルの異常によって生じる胃癌、卵巣癌、乳癌、膵癌、前立腺癌等の悪性腫瘍、糖尿病等の疾患に対する治療薬、予防薬および診断薬、脳、心臓などの虚血時における細胞死抑制剤、抗癌剤や放射線治療時における正常細胞の細胞死からの保護剤、または既存の治療薬の効果を増強する薬剤が求められている。

治療薬、<自="13"先="14D">
予防薬および<自="14"先="15D">
診断薬、<自="15"先="16D"><<自="15" 先="19D">>
脳、<自="16"先="17D">
心臓などの<自="17"先="18D">
虚血時における<自="18"先="19D">
細胞死抑制剤、<自="19"先="24D">
抗癌剤や<自="20"先="21D">
放射線治療時における<自="21"先="22D">
正常細胞の<自="22"先="23D">
細胞死からの<自="23"先="24D">
保護剤、<自="24"先="28D"><<自="24" 先="30D">>
または<自="25"先="26D">
既存の<自="26"先="27D">
治療薬の<自="27"先="28D">
効果を<自="28"先="29D">
増強する<自="29"先="30D">
薬剤が<自="30"先="31D">
求められている。<自="31"先="なし">

**図4　係り受け解析結果**

図4は例文を係り受け解析器にかけた結果である。<>の中は係り受け情報を表し、「自」はその文節の番号を表し、「先」は係り先の文節の番号を表している。また<<>>の中は人手によって求めた正しい係り受け情報となっている。この文を分割し、翻訳すると下記のようになる。

・分割翻訳結果

治療薬、予防薬および診断薬、Treatment, preventive medicine, and diagnosis medicine
脳、心臓などの虚血時における細胞死抑制剤、Cell death control medicines of brain and heart, etc. at ischemia.
抗癌剤や放射線治療時における正常細胞の細胞死からの保護剤、Protection medicine from cell death of healthy cells at anti-cancer drug and radiation therapy
またはor
既存の治療薬の効果を増強する薬剤がMedicine that reinforces effect of existing treatment

・部分訳を人手で結合したもの

Treatment, preventive medicine, and diagnosis medicine, Cell death control medicines of brain and heart, on ischemia, Protection medicine from cell death of healthy cells at anti-cancer drug and radiation therapy, or Medicine that reinforces effect of existing treatment.

・上記英文の機械翻訳結果

処置、予防医学と診断医療、脳と心臓の細胞死の制御薬（虚血の）、制ガン剤と放射線療法の健康な細胞の細胞死からの保護医療または既存の処置の影響を補強するMedicine.

この例における係り受け解析誤りは2箇所あり、1つは「診断薬」と「脳」が並列の関係になっている誤りである。この誤りは名詞＋読点による並立関係の誤りであり、正しくは「診断薬」と「細胞死抑制剤」が並立の関係にある。2つ目の誤りは「保護剤」と「効果」が並立の関係になっている誤りである。これは接続助詞「または」による並立関係による誤りであり、正しくは「保護剤」と「薬剤」が並立の関係にある。

そこで、本来並立の関係にある「診断薬」「細胞死抑制剤」「保護剤」「増強する薬剤」のところでそれぞれ分割した。その結果、この分割を行うことで「名詞＋読点」による並立の関係の誤りが回避できた。しかし、この場

合は人手で区切ったからうまくいっているのであって、後述するが、「名詞＋読点」による並立の関係は機械的に判断するのが非常に難しい。

### 4.1　分割の自動化

特許文を機械的に分割するために南瓜のオプションで形態素解析情報を表示し、人手によって分割を行ったときの分割箇所ではどうなっているのかを調査した。そして、実際に以下のような部分に注目して機械的に分割した。200文に対し分割を行い、約120文に関して、人の目で見ても自然な分割となった。

**(1)読点による分割**

読点に注目して特許文を分割する。読点を打つところは、書く人によって違うので何か文法的な規則があるわけではないが、文章を大まかに分割する上では役立つ。読点によって並立構造になっている場合、「並立助詞＋読点」なら分割しても問題ないが、「名詞＋読点」の時には、分割すると翻訳する際に意味が異なることがある。

**(2)副詞可能の名詞**

「ため」や「時」など副詞可能になる名詞に注目して分割する。このようなものを含む文節は、別に訳してからその文節が修飾する文節の後に繋げることができる。

**(3)接続助詞「て」**

「〜して」のような形になるところに注目して分割する。

### 4.2　部分間の関係

分割した文章を部分ごとに機械翻訳に通すことで係り受け解析誤りをある程度回避することができる。しかし、これはあくまで部分的なものであるので、それぞれの部分訳をつなぐ必要がある。そのためには、部分間の係り受け関係を知る必要がある。ここでは分割した場所によりそれぞれの部分がどのように関っているのか明らかになっているものについて述べる。

**(1)特許特有表現**

すでに述べたように、特許文に頻繁に現れる「本発明は、」で始まるもので、この場合「本発明は、」は文末の「構成である」などに係る。

**(2)並立助詞＋読点**

並立助詞＋読点「と、」で終わるような部分は、より下位に出てくる並立助詞＋読点の形を持つ部分、または、格助詞＋格助詞「とを」、格助詞＋連体化「との」、並立助詞＋格助詞「とに」を持つ部分に係る。

**(3)副詞句**

「ために」や「ときに」を含むような部分は、先に翻訳し、その後に続く部分の前に置いてもいいし、その後につけてもいい。

以上のような研究の結果、次のような問題点が明らかになった。

① 読点による分割

4.1では機械的に分割するための方法として読点を挙げた。読点の使い方が正しいときは問題ないが、読点を打つところは人によって違うため、読点で分割すると問題が発生する場合がある。また、「名詞＋読点」による並立関係は、機械的に判断することが難しく、正しい意味になるように分割することがうまくいかない場合が多い。

② 分割翻訳結果の結合

本研究では主に特許文の分割について調査を行い、分割翻訳後の結合についてもある程度は考察した。しかし、現状では、翻訳後の結合については調査不足であり今後もデータ収集を続け、分割後の関係を調査する必要がある。

## 5　おわりに

係り受け誤り修正システムについては、その他の分類に属するものが修正できるように現在改良中である。一つの可能性として、法令文における並列構造を利用することが考えられる。法令文においては、「および」や「ならびに」、「または」や「もしくは」という用語が用いられ、前者の方が大きな構造にかかわるとされている[14〜16]。

いくつかの例文について簡単な調査を行った結果、特許文では、必ずしもこれを守っているわけではないが、ある程度参考にできる可能性があるので、今後検討したい。

特許文の分割については、さらにデータ収集を続け、新たな分割法を提案していく必要がある。また、分割後の部分間の関係についてもより詳しく調査することにより、文章全体の大きな係り受け構造像を把握し、それぞれの部分訳をつなぎ合わせることを機械的に行うためのアルゴリズムを構築する予定である。現在考慮しているのは、動詞や形容詞の結合価を用いる[17]ということである。結合価とは、たとえば動詞「組み立てる」は、「N1がN2をN3から組み立てる」のように、3つの要素と結びつくことを示すものである。その場合に、N1, N2, N3（この場合すべて名詞）の意味を確定できれば、この動詞を含む句の範囲を定めることができる。現在結合価を用いて、どの程度動詞句が分離できるか検討中である。

最終的には、特許文を分割し、部分間の関係を求め、機械翻訳を行い、部分訳を結合することでより正確な翻訳を行うシステムの完成を目指す。

**謝辞**：法令文についてご示唆いただいたJapio奥直也氏、大塩只明氏に感謝致します。また、今野朗氏、水谷直樹氏からも適切な示唆をいただきました。ここに感謝致します。

**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：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（基盤研究（C）課題番号18500102）報告書（2007）
[3] 金田　優也：特許文の係り受け解析と分類、山形大学工学部情報科学科卒業論文（2005）
[4] 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回全国大会予稿集 6Q-3（2007）
[5] 見年代　茂大：特許文解析誤りの修正システム、山形大学大学院理工学研究科修士論文（2007）
[6] 吉田　節行：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、山形大学工学部情報科学科卒業論文（2007）
[7] 横山　晶一、見年代　茂大、吉田　節行：特許文解析誤り自動修正と、正確な翻訳のための特許文の分割、平成18年度AAMT/Japio特許翻訳研究会報告書 pp.73-83（2007）
[8] YOKOYAMA Shoichi and KANEDA Yuya: Classification of Modified Relationships in Japanese Patent Sentences, Machine Translation Summit X Workshop on Patent Translation, pp.16-20（2005）
[9] YOKOYAMA Shoichi and KENNENDAI Shigehiro: Error Correcting System for Analysis of Japanese Patent Sentences, Machine Translation Summit XI Workshop on Patent Translation（2007）
[10] 南瓜、奈良先端科学技術大学院大学
[11] Japio特許データベース（2005）
[12] 特許庁データベース http://www.ipdl.ncipi.go.jp/homepg_j.ipdl
[13] 池原　悟他：日本語語彙大系、岩波書店（1997）
[14] 田島　信威：最新法令用語の基礎知識［3訂版］、ぎょうせい（2005）
[15] 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
[16] 森　智宏：ぱてんとさいと、条文用語解説 http://patent.site.ne.jp/pa/terms.htm
[17] 荻野　孝野、小林　正博、井佐原　均：日本語動詞の結合価、三省堂（2003）