特許文書における文書構造情報の認識と活用

Recognition and Application of Document Structures for Patent Analysis

日本アイ・ビー・エム株式会社　東京基礎研究所リサーチ・スタッフ・メンバー
鈴木　祥子

2004年日本アイ･ビー･エム株式会社に入社。東京基礎研究所で数理解析のチームを経て、現在はテキスト分析のチームに所属し、特許文書や技術文書、ビジネス文書などの分析に従事している。博士（理学）。

E30126@jp.ibm.com

日本アイ・ビー・エム株式会社　東京基礎研究所主席研究員
那須川　哲哉

1989年日本アイ・ビー・エム株式会社に入社。東京基礎研究所に配属以後、IBM T.J. ワトソン研究所での１年間の勤務、コンサルティング部門への１年半の出向などを経験しつつ、一貫して機械翻訳やテキストマイニング、評判分析、会話マイニングなど自然言語処理関係の研究に従事している。博士（工学）。

nasukawa@jp.ibm.com

1　はじめに

文書には様々な構造が存在する。自然言語処理でこれまで多く研究されてきた係り受け構造はその最たる例であるが、係り受け構造以外にも、箇条書き、並列表現、及びこれらの入れ子構造、セクションなどの階層構造、文や段落の依存関係など多くの構造が挙げられる。一文内の構造に留まらず、文間の構造、文書全体中の構造を把握することは文書を理解する上で重要である。

特許文書は、上記の多種多様な構造を複雑に組み合わせて構成されている。請求項の中には、場合により多数の並列表現、またその入れ子構造が存在し、意味のある塊（構成要素）を形成している。例えば要素列挙型の記述形式では、“A と、B と、C とからなる、”のように A、B、C が並列に記述される。また、請求項間にも構造があり、これらは独立請求項・従属請求項という依存関係にある。従属請求項を理解するためにはその親にあたる請求項を理解する必要があり、権利の範囲を把握する上でも請求項の依存関係の認識は重要である。更に、明細書本文には請求項を補完するための各種記述が含まれているが、ここにも特許庁のフォーマットによって定められた構造がある。典型的な例としては、まず、【背景技術】の項目について述べた後、【発明が解決しようとする課題】の項目で従来技術における課題を述べ、【課題を解決するための手段】で発明の内容について記載する。更に場合によって、【図面の簡単な説明】が【図１】などで列挙されて記載される。図内の符号については【符号の説明】で説明されることが多い。また、多くの特許では【発明を実施するための形態】で実施形態を、【実施例】で１つから複数の実施例が記述される。

しかし、このような特許文書の構造をシステム的に把握するのは決して容易でない。請求項の要素列挙型の例の場合、A、B、C が比較的単純な名詞や名詞句であれば、並列構造の推定は容易である。しかし、これらが長い名詞句であったり複雑な並列表現を含んだ入れ子構造であったりする場合には、推定が格段に難しくなる。特許文書では新しい表現や専門用語を組み合わせた長い表現が使われることが多く、システムにとってはそういった表現が未知語になるなどして対応が困難になる。また、請求項間の依存関係抽出も、表現が多様であるため、正確に依存関係を抽出するには注意を要する。明細書本文の項目は、多くの文書で流れは共通しているものの、タグ付けが必須ではなく、表記ゆれが存在したり、項目自体が存在しない場合もある。また出願された年代によっ

てもタグの表記は異なっている可能性がある。

上記の通り、特許文書の構造を把握するのは技術的に単純ではない。しかし、このような構造を正確に把握できれば、特許情報がより有効に活用できる。本稿では、特許文書の構造解析が特許情報の活用に有用であることをいくつかの具体例を用いて紹介した上で、構造解析の手法や課題を紹介する。

2 特許文書活用における構造情報の有用性

2.1 請求項構造解析による可読性向上

特許の請求項は、記述の複雑性、および専門性の高さから、可読性が低いという問題がある。パテントクリアランスを目的として多くの特許を読む場合、請求項の可読性の低さは内容を理解する上でのボトルネックとなる。それでは、どうすれば、パテントクリアランスにおける請求項の可読性が向上するだろうか。

1つには、独立請求項はどれか、どの請求項がどの請求項に依存しているか、という情報を分かりやすく表示することが挙げられる。このためには、請求項間の依存関係を正確に抽出する必要がある。

また、請求項内の複雑な構造を木構造で表示し、発明に必要な構成要素を列挙することも効果的である。このためには、請求項内の構造を解析し、並列構造を抽出する必要がある。また、請求項を解析することで、重要語や、新規性の高い語（新規語）を推定することが可能となる[1]。ここで、重要語とは、後続の構成要素や従属請求項で繰り返し言及され意味を限定される語を指す。また、新規語とは、構成要素の依存関係および従属請求項の依存関係から推定された、発明者が新規と考える箇所に現れる語を指す。このようにして抽出された重要語や新規語の把握は複雑な請求項を短時間で理解する上で役立つ。

更に、パテントクリアランスにおいては subject matter（発明の対象）が特許選別の上で重要な情報を持っている。このため、適切な粒度の subject matter を提示できれば、不必要な請求項は読まない、などの判断が容易に出来るようになる。

また、実用上、明細書本文内の図へのスムーズなアクセスやシステムの応答性の良さも重要な要件である。

筆者らも関与した昭和電工株式会社の取り組みにおいては、これらの要件を実現し、請求項の可読性の向上を目的とした特許読解支援システムを構築、実務で実際に活用している。システムの画面例を図1に示す。

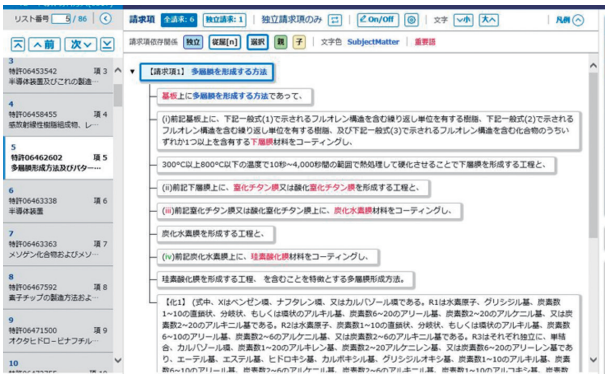

図1　昭和電工社特許読解支援システム表示例

請求項の構造解析を行い、システム上に各種機能を実装することで、独立請求項のみの表示オプション、親請求項を選択した際の従属請求項の表示（およびその逆引き表示）、各請求項の subject matter の表示、請求項内の構成要素分解、並列構造の表示、重要語・新規語のハイライト、などが実現され、一目で特許の請求の範囲を理解するために役立っている。

昭和電工社内の様々な部署でパテントクリアランスの業務の効率化を定量的に測定したところ、平均 45% の時間が削減出来たという報告が公開されている[2]。

2.2 構造解析情報による特許分析や検索の高度化

特許情報における構造抽出が、可読性向上に効果があることを前節でみてきた。次に特許分析や検索においても、構造の抽出が重要であることをみていく。

2.2.1 明細書の構造解析を通じた分析の精緻化

前節で述べたように、明細書本文には、請求項以上に情報が記載されており、それらは構造化されている。背景技術、課題、発明の実施の形態などに分かれている。大量の特許文書から有用な知見を抽出することを目的とした特許のマイニングに明細書本文の構造を利用することを考えてみよう。

ある特許集合を例にとる。この集合は IPC のメイングループが G06F 17（特定の機能に特に適合したデジタル計算またはデータ処理の装置または方法）の特許集合3万6千件（2006 年から 2019 年に出願された

もので 2020 年８月現在公開されているもの）とする。各文書内の項目（【請求の範囲】、【発明が解決しようとする課題】、【課題を解決するための手段】、【発明の効果】、【実施例】、etc.）ごとにテキストを分割し、各項目内テキスト内の形態素解析を行うことで、項目に紐づいた単語集合が得られる。【発明が解決しようとする課題】内の単語 w1 と【発明の効果】内の単語 w2 とが同一明細書内に現れるとき、w1 と w2 の関連性の強さを pointwise mutual information（PMI）という指標で算出する。PMI とは、w1 と w2 の共起のしやすさを表す統計的指標であり、w1 や w2 が出現する文書数などから導出する。このようにして多くの単語のペアについて関連性の強さの高いものを抽出すると、下記表１のようなペアの例が得られる。即ち【発明が解決しようとする課題】の中に「質問」が含まれる場合には、【発明の効果】の中に「自然だ」が含まれている傾向が強く、これにより、「“質問”に対して回答を機械が生成する技術では、回答の“自然さ”が求められる」といった知見が得られる。

表１　項目間で関連性の高い語ペアの例

【発明が解決しようとする課題】中の名詞	【発明の効果】中の形容表現	出現テキスト例
質問	自然だ	【発明が解決しようとする課題】....質問が理由や事象の説明に基づく回答を求めるNon-Factoid型質問である場合に、回答が複雑な長文になる.... 【発明の効果】....質問に対して、違和感を低減した自然な文面の回答を生成する....
撮影	鮮明だ	【発明が解決しようとする課題】....車に搭載したデジタルビデオカメラにより撮影を行ってくると,対象地域を網羅した画像データが.... 【発明の効果】....鮮明な画像だけを蓄積することができる....

この方法を模式的に表すと図２のようになる。

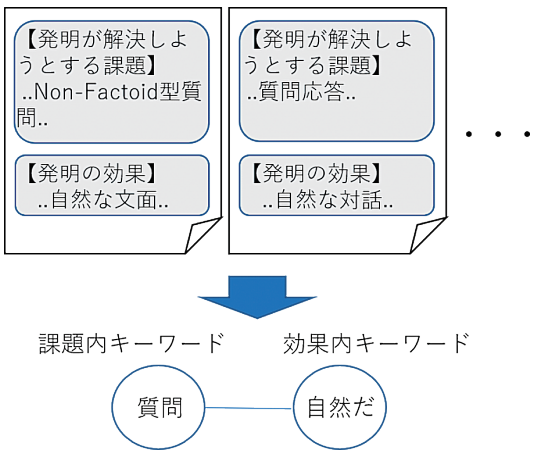

図２　項目間で関連性の高い語ペアを抽出

2.2.2　明細書の構造解析による検索の精緻化

特許検索において、ある検索クエリ（検索条件）を与えたとき、ユーザー側には当然意図がある。クエリが発明の構成要素に関するものであれば、検索対象となる文書集合中、特に【課題を解決するための手段】や【発明の実施の形態】項目などでクエリにマッチする文書を類似文書として抽出したいはずである。一方で、解決したい課題に関するキーワードをクエリとするのであれば、【発明が解決しようとする課題】項目でよりマッチする文書を類似文書として抽出したい。このようにして、クエリ自体の構造化、および検索対象の構造解析が組み合わされば、より良い検索結果が得られることが期待できる。更に、ある特許文書（クエリ文書）に対して類似特許を探す場合には、クエリ文書自体から構造化したクエリを抽出することが課題となる。この場合にも、クエリ文書の構造解析が必須となる。

ここである本願（クエリ文書）があったとき、これと似た特許文書を検索することを考えよう。例として HTML などの半構造化文書から目的とする情報を抽出して検索する発明（特開 2018-032273）を本願とする。本願には、【請求の範囲】や【課題を解決するための手段】に“半構造化文書”、“階層”、“要素”、“抽出”、“上位”、“下位”、“共通する”といったキーワードが含まれ、【技術分野】には“情報”、“抽出”、【背景技術】には“Markup”、“Language”、“HTML”、“インターネット”、“サイト”、“半構造化文書”、“階層”、“ページ”、“検索する”などが含まれる。一方、本願の拒絶理由の引用文献を類似文書として解釈し、本願と同様に項目で分解すると、【背景技術】に“Markup”、“Language”、“検索する”が含まれ、【課題を解決するための手段】に“半構造化文書”、“階層”、“要素”、“抽出”、“共通する”といった本願と共通のキーワードが含まれている。また、【発明の実施の形態】や【実施例】にもこれらの共通キーワードが出現することが分かる。このことから、本願と類似文書は項目ごとの内容も似ていると推測できる。（図３参照）

一方で、本文中に“半構造化文書”、“階層”、“抽出”、“検索する”といったキーワードが含まれている文献であっても、“半構造化文書”が【背景技術】や【発明が解決しようとする課題】のみに含まれているケースも存在する。これらの発明は、情報抽出の対象が半構造化文書に限定したものではなく、したがって半構造化文書特有の

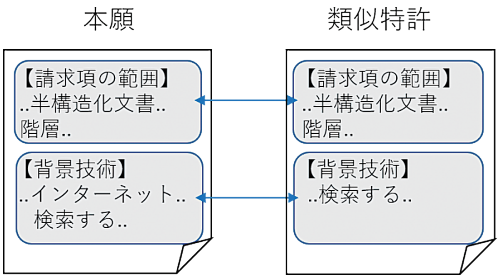

図３　項目ごと類似性による類似特許検索

特徴も利用していない。既存研究として半構造化文書の検索に触れているのみである。このように、重要なキーワードであっても出現する項目によっては意味的な類似性への影響が小さいものもある。

上記例から、項目の情報が検索に活用できる可能性をみることが出来る。ただし、前述したように、項目が必ずしも定型でないことや、項目内の記述の自由さから内容が似ていても記述が異なるケースは多く存在するため、一般の構造化データにおける項目の類似度ほどの信頼性がない点には留意が必要である。

2.2.3　明細書の構造解析による検索のクエリ拡張

前述のように、明細書本文は請求項の詳細を説明している。請求項で利用される語は抽象的な表現や専門用語が多い。このため特許を読む者は、請求項の内容を補完するために明細書本文から更に情報を得ることになる。検索においても、本願請求項から抽出された検索クエリを拡張する目的で明細書本文からもキーワードを取得するアプローチが複数提案されている[3]。この際に、補完するためのキーワードをどの項目から取得するかを選択することで、複数種類のキーワードの拡張が可能となる。

先ほどと同じ IPC=G07F 17 の特許集合を例にとる。【請求の範囲】内で“対話”というキーワードが存在する特許集合を A とする。“対話”を含む集合 A は、人間と機械との対話システムに関する発明を多く含んでいる。この A において【請求の範囲】内の“対話”と共起する単語から高い関連性を持つ単語を前述の指標 PMI で抽出すると、“発話”、“意図”、“応答”といった対話の関連表現が得られる。一方で、同じ特許集合 A において今度は【発明が解決しようとする課題】中に出現する関連の高い語として、“相槌”、“雑談”、“自然だ”、といった語が得られる。また【発明の効果】には、“共感”、“親近感”、といった語が高関連語として出現する（図４参照）。こういった関連表現は“対話”の内容をより詳細に表す語として、クエリ拡張に役立つと考えられる。項目ごとに関連語の性質が異なるため、検索したい内容によって適切な関連語を選択しクエリ拡張を行うことが出来る。

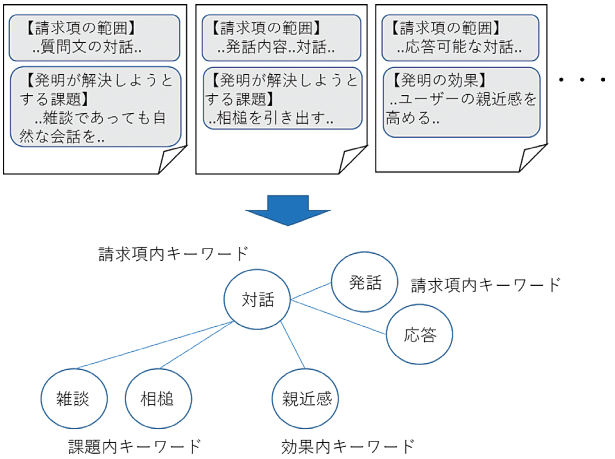

図４　項目を利用したクエリ拡張

3　特許文書構造解析のアプローチと課題

前節で、特許文書の様々な構造を利用して、可読性向上やマイニング・検索の精緻化、クエリ拡張などに活用できることを紹介した。本節では、構造解析のための手法について簡単に紹介する。同時に、構造解析の課題についても議論する。

3.1　並列構造の抽出

並列構造は、基本的に区切り文字と並列のキーとなる文字列を指定した上で、並列の範囲を求める手法が一般的である。特許文書においては各並列項目間の類似性や、並列の開始箇所の推定を手掛かりに並列の範囲を求めることが出来る。

例えば、次のような例では、並列構造が入れ子になっている。

“（A）バインダーポリマー 100 重量部当たりに、（a）（メタ）アクリル酸エステル系単量体 20 ～ 79 重量部と、（b）ビニル系単量体 20 ～ 60 重量部と、（c）不飽和カルボン酸系単量体 0.01 ～ 30 重量部と、を用いて重合されたポリマー粒子を含むバインダーと、（B）上記バインダーを用いて複数積層される電気化学セルであって、バインダーにより電極内の電極活物質粒子同士、

電極活物質と集電体とが固定及び連結されると共に、電極と上記電極と接触するセパレーターとが熱融着により接合された電気化学セルと、を備える電気化学素子であって、…”（特願 2008-511060）

ここで、まず並列のキーとなる文字列を、“と、を”と指定する。（この並列のキーとなる文字列は文書によって異なる。例えば英語の文献では、“and”および“or”を並列キーとみなすアプローチが多い。）次に、並列の区切り文字列を指定する。ここでは“と、”とする。従って、“X と、Y と、Z と、を”という構造を見つける問題になる。既存研究では、並列構造を抽出済みの正解データを教師データとして、並列構造の範囲を推定するモデルを学習するアプローチがとられていることが多い。モデルには各並列項目の類似性や可換性を利用するのが一般的である[4, 5]。

しかし大量の日本語特許文書において入れ子構造を含む正解データを人手で付与するのは手間がかかり、更に分野によっては専門性が要求されるという問題もある。そこでここでは正解データを利用しないアプローチを考える。また、特許文書においては並列項目間に依存性が存在することもあるため、並列項目の可換性は今回は利用せず、並列項目の類似性のみでどの程度並列構造が抽出できるかみてみよう。上記例文では、並列キー“と、を”が2回現れる。1回目の“と、を”は、“(a)（メタ）アクリル酸エステル系単量体 20 ～ 79 重量部”、“(b)ビニル系単量体 20 ～ 60 重量部”、“(c)不飽和カルボン酸系単量体 0.01 ～ 30 重量部”の3項目を並列に扱っている。分野の特色上、これらの3つの項目は文字列的にも非常に類似している。2回目に出現する並列キー“と、を”は、より大きな並列構造を持ち、“(A) バインダーポリマー 100 重量部当たり…ポリマー粒子を含むバインダー”と、“(B) 上記バインダーを用いて…接合された電気化学セル”とが並列に列挙されている。これらもまた、“(A)”と“(B)”から始まる記述に類似性がある。このようにして並列に並ぶ各アイテムの類似性が判定できれば、入れ子構造を含む並列構造の抽出へのアプローチが可能となる。

また、この例のように箇条書きになっている場合には、{(A), (B), …} や {(a), (b), (c), …} という表現が列挙を表すという知識を利用すれば、並列構造の抽出精度は高まると期待できる。

しかしながら、複雑な入れ子構造をとる並列構造抽出は、未だチャレンジングなタスクである。特に特許文書からの並列構造抽出は、専門分野ごとに記述スタイルが異なり、また膨大な教師データが存在しないため、様々な工夫が必要である。

3.2 明細書本文の構造の抽出の課題

前節の並列構造抽出と比較すると、明細書本文の項目の抽出は簡単なタスクと感じる読者が多いかもしれない。実際、明細書本文は記述ルールが存在するため、項目ラベルの表記ゆれをまとめて対応関係を作成すれば項目の抽出が可能である。しかし、出願年代によって項目の傾向が変わることもあり、項目の対応関係の作成やメンテナンスは課題であり続ける。また、思いがけない項目ラベルの記述があった場合には対応できないという問題も発生する。

さらには、明細書本文の各段落間にも、依存関係などの複雑な関係が存在する。単純な例としては、実験手順が順序つきリストで記述されているケースが考えられる。あるいは請求項の構成要素間依存関係に類似した依存関係が段落間に存在するケースも非常に多い。

明細書本文のリッチな情報を生かすためには、存在する構造を可能な限り抽出し、またこれらの構造を利用した分析が必要になると考える。

4 まとめ

本稿では、特許文書を例に、並列構造、依存関係、項目による分割など文書内の様々な構造について解説した。また、特許文書の構造を利用することで、可読性の向上、特許分析や検索の高度化など様々なメリットが得られる可能性を指摘した。

このような構造は、特許文書に留まらず、論文、製品マニュアル、社内文書、法的文書などあらゆる文書が普遍的に持つ構造である。構造ルールを正確に把握することが出来れば、各文書から目的となる情報を効率よく抽出することが可能となる。構造を持った文書に対する機械によるアプローチは、これまで XML など機械的な処理を前提として、明確に定義された構造ルールで記載されたものが中心であった。一方で、特許文書のような、機械処理を第一の前提としない文書には、ルールや

記述に揺らぎがある。またルールが明示的に定められていないことも多く、まずルールの把握から行う必要があるケースも多い。このような文書から正しい構造を抽出し、分析に生かしていくことが重要であると筆者らは考える。

* 本稿記載の内容は筆者個人の見解に基づいています。

参考文献

[1] Shoko Suzuki and Hiromichi Takatsuka, “Extraction of Keywords of Novelties from Patent Claims,” Coling, 2016.

[2] 昭和電工株式会社ニュースリリース https://www.sdk.co.jp/news/2019/27188.html

[3] 奥村　学, “特許情報処理：言語処理的アプローチ”, コロナ社, 2012.

[4] Jessica Ficler and Yoav Goldberg. “A Neural Network for Coordination Boundary Prediction”, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, p.23-32, 2016.

[5] Hiroki Teranishi, Hiroyuki Shindo, and Yuji Matsumoto, “Decomposed Local Models for Coordinate Structure Parsing”, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), p.3394-3403, 2019.