特許明細書からの出願目的・技術課題情報の抽出

Extraction of Information on the Purpose and the Technical Effect from a Patent Document

酒井 浩之[*1] 野中 尋史[*1] 増山 繁[*1]
Hiroyuki Sakai Hirohumi Nonaka Shigeru Masuyama

[*1]豊橋技術科学大学
Toyohashi University of Technology

We propose a method for extracting information on the purpose and the technical effect from a patent document. The information on the technical effect extracted by our method is useful for generating patent maps (see e.g., Figure 1) automatically or analyzing the technical trend from patent documents. Our method extracts expressions containing the information on the technical effect by using clue expressions effective for extracting them. Our method achieves not only high precision (77.1%) but also high recall (76.3%) by acquiring such clue expressions automatically from patent documents.

1. はじめに

現在，新規の特許出願は年間４０万件以上といわれ，効率的に特許分析を行うための技術が求められている．パテントマップは特許の出願を可視化したものであり，企業における技術開発戦略，及び，知財戦略の策定や国，地方自治体における技術開発推進政策立案に使用される．ここで特許庁作成のパテントマップ[*1]に記載されているような技術課題（特許発明を使用することにより解決される課題），解決手段（特許発明の構成要素）を軸として，特許出願動向を可視化したもの（図１を参照）は，下記に示すように効率的な特許分析を行ううえで特に有益である．

- 技術課題は，該当する発明が技術の利用者に与える便益，すなわち利用者のニーズを示し，解決手段は，技術の詳細な内容を要約したもの，すなわち技術シーズに相当する．よって，上記パテントマップは，競合企業における特定のニーズに対応する技術シーズの内容を把握できること，及び，ニーズとシーズ両方についてパテントポートフォリオにおける自社と他社の強み・弱味の分析を容易に行えることから，自社の技術開発戦略策定に役立つ．

- 特許庁における特許の審査が技術課題と解決手段の両方を加味して行われることもあり，知財戦略策定上必要となる自社特許と類似の技術課題・解決手段を持つ競合特許群の把握に，上記パテントマップは大きく寄与する．

- 上記パテントマップを使用すれば，国家等が策定している技術開発の指針となる技術戦略マップ（通常は，ニーズとシーズ，双方に着目した内容で構成される）等と比較した現時点の民間企業・大学等における技術開発状況を容易に把握できる．また，重点分野にも関わらず技術開発が遅れている分野の特定が可能となるため，そのような分野の技術開発を重点的に促進する政策立案を促す効果があるなど，国家等の政策立案にも有用な情報を提供する．

連絡先: 豊橋技術科学大学，豊橋市天伯町雲雀ヶ丘 1-1，0532-44-6867，0532-44-6873，sakai@smlab.tutkie.tut.ac.jp

*1 http://www.jpo.go.jp/shiryou/s_sonota/tokumap.htm

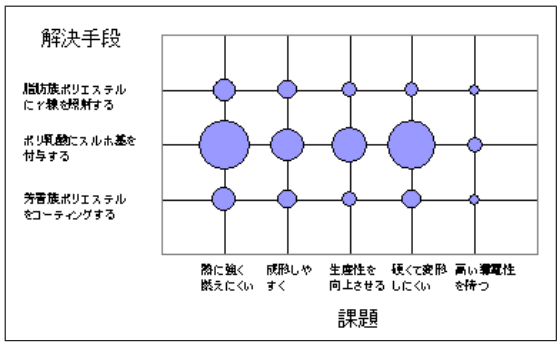

図 1: パテントマップの例

このように，パテントマップを作成することで，効率的，かつ，有効な特許分析が可能になるが，現状ではパテントマップの作成は特許庁等が専門家を利用して主に手動で行っており，多大なコストがかかる．そこで，本研究では，パテントマップ自動作成のための情報源として，このうちの技術課題の抽出に着目し研究を行う．また，特許明細書における出願目的の情報を分析することで，技術トレンドの時系列的な推移などが分かり，重点技術分野の特定や技術ロードマップ作成の素材としても有用である．そこで，本研究では，特許明細書から出願目的の情報を含む文を抽出することも行う．具体的には，「信号処理装置を提供」といった出願目的や「３Ｄ画像を容易に作成することができる」といった出願者が解決しようとしている技術課題に相当する情報を，「ができる。」のような技術課題情報を抽出するのに有効な手がかり表現を自動的に獲得し，それを使用することで特許明細書から抽出する手法を提案する．

2. 関連研究

パテントマップ自動生成の関連研究として，Uchida らは特許明細書をクラスタリングしたのち，各クラスタに適切なラベルを付与することでパテントマップを生成する手法を提案している [Uchida 04]．ただし，付与するラベルはいくつかのラベル候補であるキーワードから人手で選択しており，「技術課題」や「解決手段」に関連したキーワードを自動的に付与しているわけではない．それに対して，本研究で抽出する技術課題情報に含まれるキーワードを取得することで，「技術課題」に関連したキーワードを取得可能である．また，技術課題情報に含ま

れる表現は，図 1 のようなパテントマップを作成するために必要な情報となる．ただし，特許明細書から図 1 のようなパテントマップを自動的に作成するには，特許明細書から技術課題情報を抽出したのち，技術課題情報を含む文から「解決手段」に相当する情報を獲得する技術が必要となるため，本手法だけではパテントマップを自動的に作成することはできない．そのため，本手法によって抽出された技術課題情報を含む文から「解決手段」に相当する情報を獲得し，自動的にパテントマップを作成する手法は今後の課題とする．

新森らは，特許明細書の「発明の効果」「課題を解決するための手段」「発明の実施の形態」と，特許請求項とを自動的に対応付ける手法を提案している [新森 05]．そして，対応付けされた文の特許請求項と対応付けられた箇所より後続する箇所から，特許請求項の効果に該当する情報を得ることができ，このような情報は本研究で対象としている技術課題情報と同一のものである．しかし，特許請求項の内容を含まず，効果だけを述べた文も存在し，そのような場合は特許請求項の効果に該当する情報を得ることができない．それに対して，本手法では技術課題情報を抽出するために有効な手がかり表現を使用して，特許明細書の技術課題情報を抽出するものであり，発明の効果だけを述べた文からも技術課題情報を抽出できる．

石川らは,「ことにより」という表現を手がかり表現として，特許明細書から手段とその効果から構成される因果関係知識を抽出する手法を提案している [石川 04]．本手法においても，技術課題情報の抽出に有効な手がかり表現を使用することで技術課題情報の抽出を行う．しかし，本手法での手がかり表現は，最初に与える 1 つの初期手がかり表現「ができる。」から自動的に獲得されるものであり，人手で用意したものではない．そのため，最初に与える初期手がかり表現を変更することで，本手法を特許明細書からの他の情報（例えば,「解決手段」情報など）を抽出するタスクへ適用することが可能である．また，石川らの手法では,「ことにより」を使用していない効果のみを述べた文からは因果関係を抽出できないが，本手法で自動的に獲得される手がかり表現は，発明の効果だけを述べた文からも技術課題情報を抽出できるものである．

3. 出願目的情報の抽出

本手法では，出願目的情報を，特許明細書の「発明が解決しようとする課題」タグに該当する文集合から取得する．本手法では,「発明が解決しようとする課題」タグに該当する文集合のうち,「本発明は」「を提供」「を課題」「を目的」のいずれかを含む文を出願目的情報として抽出した．以下に，抽出した出願目的の例を示す．

> そこで、本発明は、転落事故後の迅速な事故処置を図るために、転落事故を適切に検知し、また、転落事故の発生を通報することを目的とする。

4. 技術課題情報の抽出

本手法では，技術課題情報を「発明の効果」タグに該当する文集合から取得する．具体的には，文集合から「ができる。」といった技術課題情報を抽出するための手がかりとなる表現（以降,「手がかり表現」と定義）を使用することで，技術課題情報を抽出する．以下に，特許明細書における「発明の効果」タグに該当する文の例を示す．この例では，2 つある太字の部分が抽出すべき技術課題情報となる．

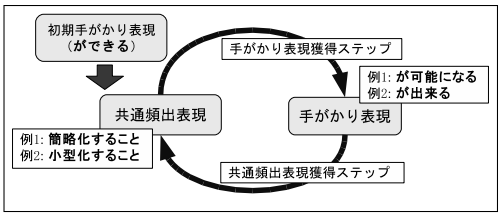

図 2: 手がかり表現自動獲得手法の概要

> 請求項２に記載の発明の現像装置によれば、請求項１に記載の発明の効果に加え、**ワンウェイクラッチのためのスペースと強度を確保できる**とともに、**現像ユニットを小型で軽量にすることができる。**

しかしながら，技術課題情報を抽出するために有効な手がかり表現の種類は数多いため，全ての手がかり表現を網羅したリストを作成することは困難である．更に，人手で作成した少数の種類の手がかり表現を技術課題情報の抽出に使用することは，再現率の低下の原因となる．そのため，本研究では，手がかり表現を「発明の効果」タグに該当する文集合から自動的に獲得し，それらを使用することで上記の情報を抽出する．

4.1 手がかり表現の自動獲得手法の概要

本節では,「ができる。」のような技術課題情報を抽出するのに有効な手がかり表現を,「発明の効果」タグに該当する文集合から獲得する手法について説明する．なお，以前，我々は，交通事故事例記事から「前方不注意が原因」のような交通事故原因表現，業績発表記事から「電子部品の売り上げが好調」のような業績要因表現を抽出するための手法を開発した [Sakai 08][酒井 06]．本手法は，それらを特許明細書から技術課題情報を抽出するために適用し，いくつかの規則を加えて改良したものである．

まず，本手法の概要を以下に示し，4.2 節，4.3 節で各処理について詳細に説明する．

Step 1: 1 つの手がかり表現（「ができる。」）を人手で与える．

Step 2: 手がかり表現の直前に出現する可能性が高い表現を取得する．（以降，手がかり表現の直前に出現する可能性が高い表現を共通頻出表現と定義する．例えば「簡略化すること」といった表現が抽出される．詳細は 4.2 節で述べる．）

Step 3: 共通頻出表現から，新たな手がかり表現を獲得する．（詳細は 4.3 節で述べる．）

Step 4: 獲得した手がかり表現から，新たな共通頻出表現を獲得する．（Step 2 と同一の処理）

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなるか，もしくは，予め定めた回数まで繰り返す（図 2 を参照）． □

4.2 共通頻出表現の獲得

本節では，共通頻出表現の自動獲得について述べる．まず，共通頻出表現の候補を取得する．ここで，図 3 のように,「こと」+手がかり表現を含む文を抽出し，その文の「こと」を含む文節に係る文節に,「こと」を追加した表現を共通頻出表現候補とする．例えば,「量産すること」のような表現が共通頻出表現候補として取得される．ただし,「すること」「させること」を共通頻出表現候補から除外した．

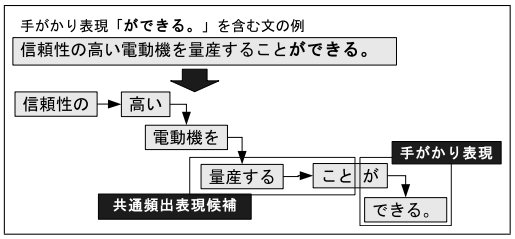

図 3: 共通頻出表現候補の取得

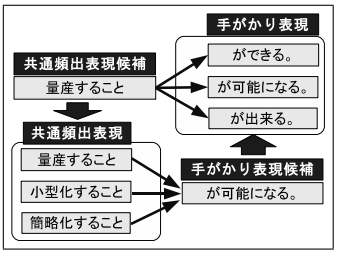

図 4: 共通頻出表現・手がかり表現の選別

次に，共通頻出表現候補の中から，適切な共通頻出表現を選別する．具体的には，図 4 のように様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを式 1 で求め，その値が閾値 T_e 以上の共通頻出表現を選別する．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \quad (1)$$

ただし，$P(e,s)$ は「発明の効果」に該当する文集合において，共通頻出表現 e が手がかり表現 s に係る確率，$S(e)$ は共通頻出表現 e が係る手がかり表現の集合である．閾値 T_e は，以下の式 2 によって設定する．

$$T_e = \alpha \log_2 |Ns| \quad (2)$$

ここで，Ns は共通頻出表現を取得するのに使用した手がかり表現の集合，α は定数 $(0 < \alpha < 1)$ である．ただし，初回は手がかり表現の数が初期手がかり表現「ができる。」の 1 つなので，共通頻出表現のエントロピー $H(e)$，および，閾値 T_e が 0 になる．そのため，初回のみ「ができる。」から取得される全ての共通頻出表現候補が共通頻出表現として選別される．

4.3 共通頻出表現を使用した手がかり表現の獲得

共通頻出表現の選別を行った後，その選別した共通頻出表現から新たな手がかり表現を獲得する．まず，抽出した共通頻出表現を含む文を抽出し，その中で共通頻出表現を含む節 P_a が係っている文節 P_b を獲得する．次に，P_a に含まれる格助詞を P_b に追加し，それを手がかり表現候補とする．例えば,「が可能になる。」のような表現が手がかり表現候補として取得される．ただし，追加する格助詞が「に」であった場合，手がかり表現候補としない．これは，手がかり表現候補として,「により、」「になり、」などの原因を表す箇所に後続する表現が取得されるためである．また，P_b が「ある」,「し、」である場合，および，P_b が「している」,「した」,「される」,「された」,「ない」,「なく」,「する」を含む場合は，手がかり表現候補から除外した．ここで，図 4 のように，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定に基づき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを式 3 で求め，閾値以上の候補を手がかり表現として抽出する．

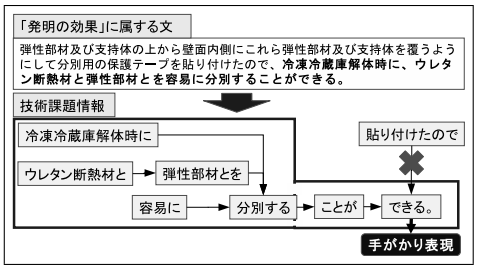

図 5: 技術課題情報の抽出 1

$$H(s) = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e) \quad (3)$$

ただし，$P(s,e)$ は「発明の効果」に該当する文集合において，手がかり表現 s に対して共通頻出表現 e が係る確率，$E(s)$ は手がかり表現 s に係る共通頻出表現の集合である．閾値は，共通頻出表現と同様に式 2 によって設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．本手法によって獲得された手がかり表現を技術課題情報を抽出するために使用する場合，手がかり表現に追加した格助詞を除去した文字列となる．表 1 に，閾値を決定するためのパラメータ α を 0.5 とした場合に，本手法によって獲得された手がかり表現を全て示す[*2]．この結果は，358,085 件の特許明細書を国際特許分類のセクション（A〜H）に基づいて分類された各特許明細書集合から上記の手法を用いて手がかり表現を獲得し，5 つ以上の特許明細書集合から獲得された手がかり表現である．これは，共通頻出表現は特許明細書集合の内容によって変化するが,「ができる」のような適切な手がかり表現は内容によって変化せず，したがって，複数の特許明細書集合から獲得された手がかり表現は適切であるという仮定に基づく．

4.4 手がかり表現を使用した技術課題情報の抽出

獲得した手がかり表現を使用して技術課題情報を抽出する手法を以下に示す．

Step 1: 手がかり表現を含む文を抽出し，手がかり表現を含む文節を検索する．(以降，手がかり表現を含む文節を P_{clue} とする．)

Step 2: 図 5 のように，P_{clue} の直前に出現し，P_{clue} に係っている文節を P_{clue} の直前に追加し，それを P_{clue} と再定義する．その後，係り元がなくなるまで，P_{clue} に係っている文節を P_{clue} の直前に連接していき，技術課題情報として抽出する．

Step 3: P_{clue} に読点が含まれる場合，Step 2 の処理だけでなく，図 6 のように，P_{clue} が係っている文節も P_{clue} の直後に追加して，技術課題情報として抽出する．

ただし，P_{clue} に係っている文節が「さらに」のような副詞であった場合，その前に出現する文節を P_{clue} に追加する．

5. 評価

本手法を実装して評価を行った．実装にあたり，形態素解析器として ChaSen[*3]，係り受け解析器として CaboCha[*4]を使

*2 表 1 の手がかり表現は人手による選別を行っていない．
*3 http://chasen-legacy.sourceforge.jp/
*4 http://chasen.org/~taku/software/cabocha/

表 1: 獲得された手がかり表現（追加した格助詞を除去したもの）

できるので	できるとともに、	でき、また	できて、	可能になり、	可能と	できるようになる。	出来る	出来、	
でき、かつ	できると共に、	できるとともに	出来るので、	できるようになり、	出来る。	できると共に	できるから、		
でき	できる、	可能であり、	できた。	可能で、	できるし、	できる。	でき、より	でき、かつ、	できる
できるようになった。	できるという	可能な	できるので、	できて	でき、	可能である。	でき、また、	可能になる。	

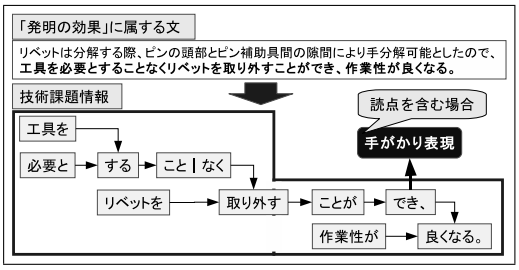

図 6: 技術課題情報の抽出 2

表 2: 評価結果

	手がかり表現	精度 (%)	再現率 (%)
本手法	36 個	77.1	76.3
ベースライン	1 個	83.7	48.6

用した．特許明細書として 358,085 件を使用し，本手法によって技術課題情報を抽出した．なお，本手法における閾値決定のためのパラメータ α は 0.5 とし，共通頻出表現獲得，および，手がかり表現獲得の繰り返し回数を 5 回とした．評価用の正解データは，無作為に 82 個の特許明細書を選び，その中の「発明の効果」タグに該当する文集合から技術課題情報を含む表現を人手で抽出して作成した．そして，本手法によって抽出された技術課題の表現が，正解として抽出された技術課題の表現を含む場合を正解として，精度，再現率を計算した．ここで，本手法に最初に与える手がかり表現である「ができる。」のみで技術課題情報を抽出する場合をベースラインとした．評価結果を表 2 に示す．

6. 考察

表 2 より，ベースライン手法における手がかり表現「ができる。」のみでは精度は 83.7% と高いものの，再現率が 48.6% であり，本来，技術課題として抽出すべき情報の半分しか抽出できないことが分かる．それに対して，本手法では「ができる。」から自動的に獲得した手がかり表現により，再現率が 76.3% まで上昇した．また，本手法の精度は 77.1% であるため，手がかり表現が増えたことによりベースライン手法に比べて低下したものの，低下は 6% 程度であった．この結果から本手法は技術課題情報を抽出するのに有効な手がかり表現を自動的に取得できたと考える．

本手法で間違って取得したもの，取得できなかったものを以下に挙げる．

間違って抽出した例：請求項 1 に記載の構成と同様な効果を奏する画像形成装置を提供することができる。
抽出できなかった例：良好な画像印刷が安定して得られる。

間違って取得した例では，技術課題情報として無価値なものを抽出した．本手法では，文の内容を考慮せず，手がかり表現を含んでいれば技術課題情報として抽出している．そのため，「発明の効果」に属する文だけでなく，特許明細書の「請求項」等の情報を使用し，その文との類似度を測ることで，例のような表現が技術課題情報として抽出されることを防ぐことができると考える．取得できなかった例では，本手法で獲得した手がかり表現を含んでいなかったため，抽出できなかった．ただし，共通頻出表現として「得ること」を獲得しており，「得ること」から「得られる」を生成することが可能である．同様にして共通頻出表現の「動詞＋こと」を「られる」の形式に変換することで手がかり表現を自動生成した．その結果，手がかり表現の総数は 312 個となり，精度 75.8%，再現率 77.7% となった．このように，再現率は上昇したが精度が低下し，技術課題情報に頻出する動詞の選別を行う必要があると考える．

7. まとめ

本研究では，特許明細書から出願目的・技術課題情報を自動的に抽出するための手法を提案した．本手法によって抽出された技術課題情報は，図 1 のようなパテントマップ自動生成や，技術トレンドの分析に有用である．本手法では，技術課題情報を抽出するために有効な手がかり表現を使用して抽出するが，この手がかり表現を自動的に獲得することで，高い精度 (77.1%) だけでなく高い再現率 (76.3%) をも達成することができた．今後の課題として，本手法によって抽出された技術課題が含まれる文から「解決手段」の情報を抽出し，パテントマップを自動生成する手法の開発を挙げる．しかしながら，「解決手段」情報の抽出においても手がかり表現が有効であり，その自動獲得のために本手法が適用できると考える．

参考文献

[石川 04] 石川 大介, 石塚 英弘, 宇陀 則彦, 藤原 譲.: 特許文献における因果関係の抽出と統合, 情報知識学会誌, Vol. 14, No. 4, pp. 105–118 (2004)

[酒井 06] 酒井 浩之, 梅村 祥之, 増山 繁.: 交通事故事例に含まれる事故原因表現の新聞記事からの抽出, 自然言語処理, Vol. 13, No. 4, pp. 99–124 (2006)

[Sakai 08] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008)

[新森 05] 新森 昭宏, 奥村 学.: 特許請求項読解支援のための「発明の詳細な説明」との自動対応付け, 自然言語処理, Vol. 12, No. 3, pp. 111–128 (2005)

[Uchida 04] Uchida, H., Mano, A., and Yukawa, T.: Patent Map Generation using Concept-based Vector Space Model, in *Working Notes of NTICR-4* (2004)