

効果—技術型パテントマップ生成を目的とした特許明細書中からの効果語 及び技術語の自動抽出と意味的統合

Extracting Effect-Technology Terms and Semantic Processing for Generation of an Effect-Technology Type Patent Map

野中尋史^{*1,2} 小林暁雄^{*1} 坂地泰紀^{*1} 鈴木祐輔^{*3} 酒井浩之^{*4} 増山 繁^{*4}
Hirofumi Nonaka Akio Kobayashi Hiroki Sakaji Yusuke Suzuki Hiroyuki Sakai Shigeru Masuyama

^{*1} 豊橋技術科学大学大学院電子・情報工学専攻 ^{*2} 豊橋技術科学大学産学連携推進本部

^{*3} 豊橋技術科学大学大学院知識情報工学専攻 ^{*4} 豊橋技術科学大学大学院情報・知能工学専攻

A patent map, a visual representation of related patent information, is an effective strategic tool for analysis of patent application trends. In particular, an effect-technology type patent map is commonly used as patent examinations are based on technology and effect terms of the inventions. However, conventional patent mining tools are not realized to make an effect-technology type patent map. In order to generate effect-technology type patent maps automatically, we propose a method for extracting information on the effect and the technological terms from patent documents. Our extraction method for effect terms achieves high precision (85.0%) by using frequent expressions, clue expressions, and the grammar patterns. Moreover, our semantic processing method attains high precision (93.0%) by using clustering method.

1. はじめに

パテントマップとは、技術分野毎に特許の出願動向を可視化したものを言う。パテントマップには、いくつかの種類があるが、特に、「発明の効果」と「解決手段」の二つを軸として表現したパテントマップ(以下、「効果-技術型パテントマップ」と呼ぶ)は、知財戦略や研究戦略の策定などで重要な役割を果たす。効果-技術型パテントマップが特に重要である理由は、特許庁の審査官が特許を審査する際、「発明の効果」を加味しながら「解決手段」に相当する技術内容を精査すること、および、特許中の課題・発明の効果(以下、効果表現と呼ぶ)は、技術が生み出す産業上のメリット(たとえば、コスト削減など)を意味し、解決手段は、発明の構成要素、すなわち、技術内容を表し、特許文書の内容を最大限縮約すれば、効果と技術内容の2つに集約されるためである。

例として、特許庁が作成したアルミリサイクルに関する効果-技術型パテントマップを図1に示す。図1のような効果-技術型パテントマップは、現状では手作業で作成されている。しかし、多大な作業時間を要する上に、特許に関する知識と技術分野に関する知見の両方を兼ね備えた専門家が作業を行わねばならないため、多大なコストがかかる。

一方、従来より、テキストマイニングを利用したパテントマップ自動作成に関する研究[Uchida 2004]等が提案されている。しかしながら、上記の研究をはじめとする従来研究は、語の文書内に出現する頻度と出現する文書数の逆数である $tf-idf$ をはじめとする重みづけを語に行った上で、語を基底とするベクトル(効果や技術内容に分類することは行わない)を生成し、そのベクトルに対して、主成分分析等を施し次元数を縮減(多くは2次元)しマップ化するものである。このため、図1のような効果-技術型パテントマップのように、効果と解決手段により特許を分類し、自動的にマップ化するものではない。

図1のような効果-技術型パテントマップを生成するためには、特許文書中に出現する語のうち、効果に相当する語(例:「寿命向上」)と技術内容に相当する語(例:「プレス圧搾」)を抽出する必要がある。

効果-技術型パテントマップを自動生成する基礎となる従来研究として、効果や特長に相当する表現(以下、効果表現と呼ぶ)を抽出する研究[石川 2004], [西山 2008]がある。しかし、これらの手法の問題点として、一部の係り受けに使用される表現(「ことにより」や「改善する」等)を辞書に定義し、抽出するものであったので、網羅性に欠けるなどの欠点があった。この欠点を解消すべく、表現を自動的、かつ網羅的に抽出する手法として、酒井らの手法[酒井 2009]がある。

ただし、酒井らの研究も含めて抽出対象としているのは、語レベルではなく、表現レベルでの抽出であり、そのままではパテントマップを自動生成することではでない。また、技術をうまく抽出できたとしても、それらの語の間には意味的に重複するものが存在する。例えば、「機械強度」と「機械的強度」は同じ意味を示すものであり、パテントマップを生成する上では、これら重複する意味を持つ語を統合する必要がある。しかしながら、従来の研究では、意味的な重複を持つ語の取り扱いについて考慮したものは知る限りでは存在しない。このため、図1のような効果-技術型パテントマップを自動的に作成するためには、特許文書中の項目の意味レベルでの解析を行い、その効果に相当する要素と技術に相当する要素を抽出し、マッピング化する手法を確立する必要がある。

そこで、本研究では、効果に相当する語(効果語と呼ぶ)の抽出に焦点をあて、自動的かつ網羅的に効果表現を抽出できる酒井らの手法[酒井 2009]に基づき、その表現中から、ノイズを除去し、直接的に効果を表す語(以下、効果語と言う)を抽出することで、効果-技術型パテントマップの軸のうち効果に相当する語を抽出し、さらに、意味的に重複する語を統合する手法の基礎的検討を行った。以下、手法とその評価結果について詳しく説明を行う。

連絡先:野中尋史, 豊橋技術科学大学, 愛知県天伯町雲雀ヶ丘1-1, 0532-44-6867, nonaka@smlab.tutkie.tut.ac.jp

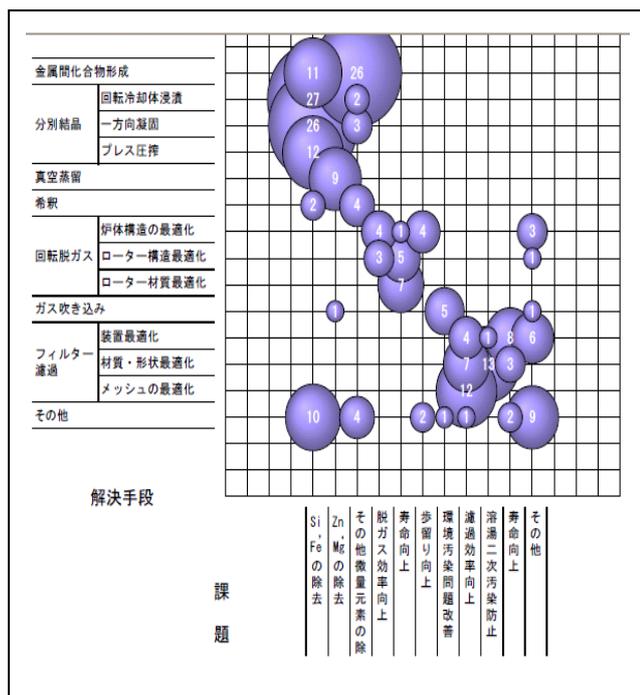


図1. 効果-技術型パテントマップの例

2. 提案手法の概要

提案手法は、大きく、効果語(一部の文書では技術語も)を抽出する部分と意味的な統合を行う部分に分かれる。

効果語抽出部では、酒井らの手法[3]で効果表現を抽出した上で、文法パターンを利用し、効果語(図1では、「機械特性」)抽出を行う。各特許明細書より効果語を抽出した後、語の形態素を素性としたクラスタリングを行うことで重複した意味を持つ語の統合を行う。これらの処理により効果-技術型パテントマップを自動的に生成することが可能となる。

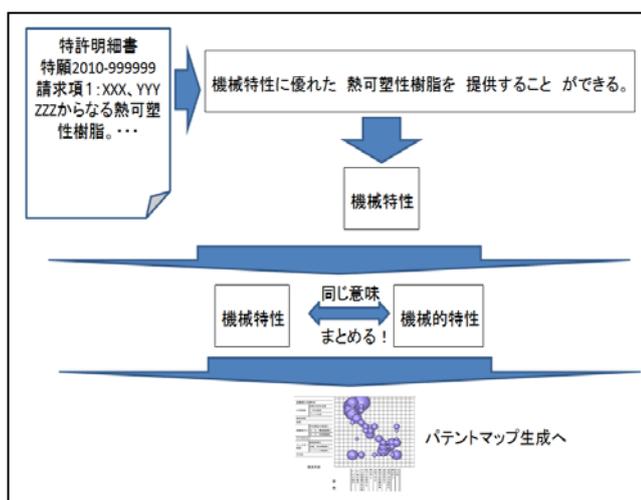


図2. 提案手法の概略

3. 効果表現の抽出

効果語を抽出する前処理として、効果語が含まれる表現を特許中から抽出しておく必要がある。

効果表現の抽出については、酒井らの手法[5]を用いる。酒井らの手法は、図3に示すように、初期手がかり表現とと呼ばれる1つの手がかり表現(「ができる。」)を手で与えた上で、手がかり表現の直前に出現する可能性が高い表現(以下、共通頻出表現と呼ぶ)を取得し、さらに、新しい手がかり表現を抽出するという作業を繰り返し行う(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)ことで、網羅的に効果に相当する表現の手がかりとなる表現を抽出する。共通頻出表現取得は、式(1)に示すようにエントロピー $H(e)$ を用いて行い、閾値以上のものを選択する。

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

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。手がかり表現の抽出も同様に行う。

本手法により、手がかり表現としては、「でき」、「できる」、「可能」などの26個が取得された。

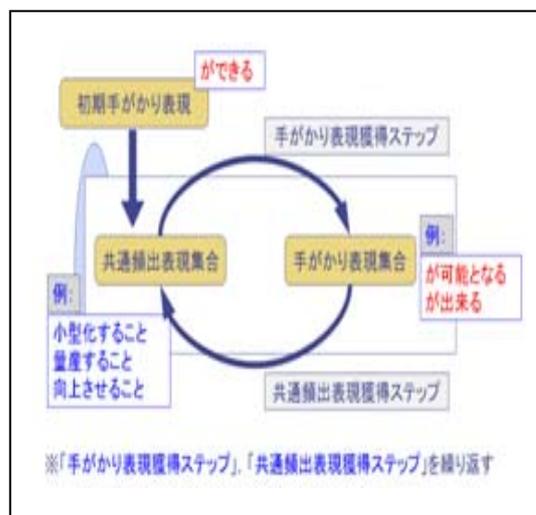


図3. 効果表現取得手法の概要

4. 効果語の抽出

効果表現として取得されるものの例を、以下、例1から例5に示す。

- 例1:「機械的性質をさらに向上させることができる。」
- 例2:「多量に廃棄されている中質繊維板粉や樹脂の廃材を再利用して、引張強度、曲げ強度等の機械的強度を改善し、かつ、木材の風合いを有する組成品を得ることができる。」
- 例3:「廃棄時の処理やリサイクルを容易化できる」
- 例4:「耐熱性、耐久性にも優れた熱可塑性樹脂組成物を、低コストで得ることができる。」
- 例5:「低コストで製造することができる。」

効果語の抽出ステップでは、
 例1, 例2, 例4は「機械特性」、例3は「リサイクルの容易化」
 例4, 例5は「低コスト(コスト低減)」, のように、表現中に現れる効果に直接相当する語の抽出を行う。

本研究では、予備調査に基づき同定した、以下、図4から図6に示す文法パターン1から文法パターン3に示されるような、酒井らの手法[酒井 2009]で取得される共通頻出表現、手がかり表現と効果語の抽象的な文法関係より効果語抽出を行う。

ここで、パターン1は、効果語が直接、手がかり表現にかかる(共通頻出表現がそのまま効果語になる)ものである。例えば、「防音する」のようなものである。

また、パターン2は、効果語を目的語とし、共通頻出表現、手がかり表現とつながるものである。例えば、「機械強度を向上することができる(「機械強度」が効果語、「を向上する」は共通頻出表現)、「できる」は手がかり表現)」のようなものである。

さらに、パターン3は、パターン2の変形で、効果語が特定の手がかり表現(以下、効果語手がかり表現と呼ぶ、「優れ」等)を通じて、技術語表現に接続するものである。例えば、「機械特性に優れた熱可塑性樹脂を提供することができる。」のようなものである。

ここで、各パターンのどれに属するかを決定するためには、技術語の判定が必要である。技術語の判定は、特許請求項と表現中の複合名詞とのマッチングにより行う。これは、特許請求項には技術語しか含まれないという特性を利用したものである。なお、請求項に含まれない技術語も存在するため、請求項とのマッチングにより得られる語がよく係る共通頻出表現(提供する)をDF値により検出し、該当共通頻出表現を用いて、その表現に係る語を技術語と判定することにより、網羅的に技術語を取得する。

上記より、技術語を判定し、図7に示すステップに基づきパターンを識別し、効果語を抽出する。

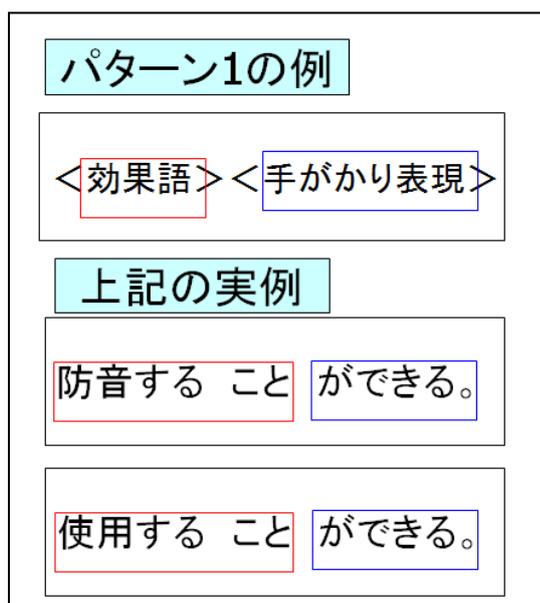


図4. パターン1

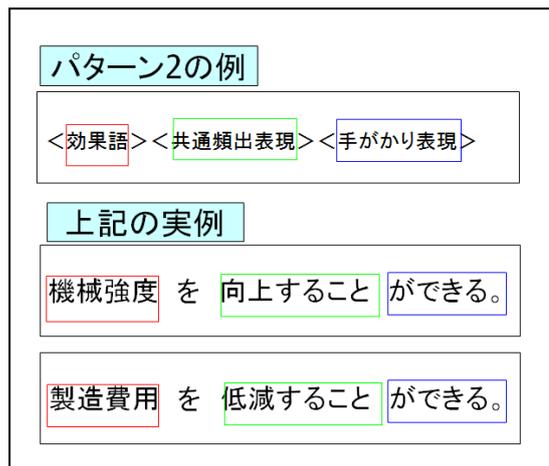


図5. パターン2

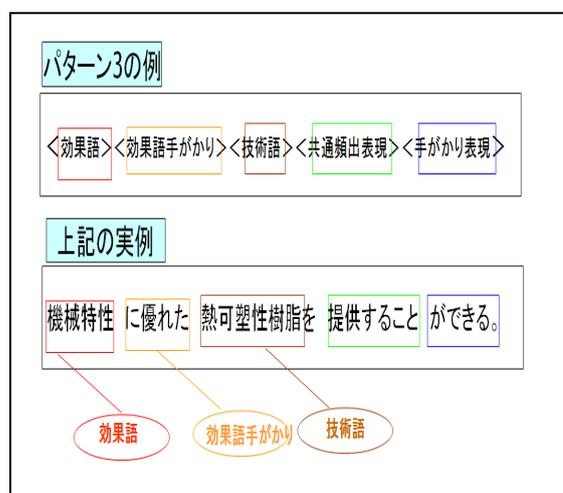


図6. パターン3

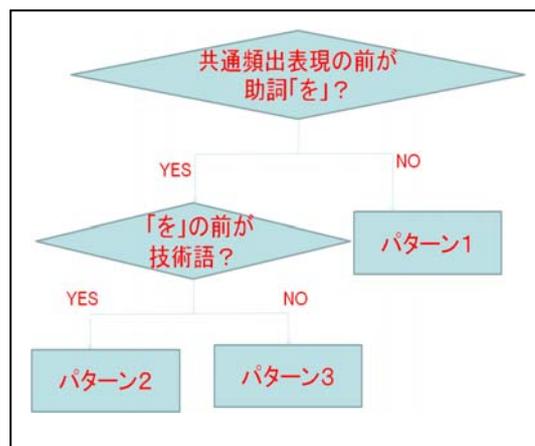


図7. 文法パターン判別のフローチャート

5. 効果語の意味的統合と実験結果

「機械特性」と「機械的性質」や、「低コスト」と「コスト低減」等のように、重複した意味を持つ効果語が存在する。このため、パターンマップの軸の要素として効果語を使用する場合、重複した意味を持つ語は、ひとつにまとめあげ必要がある。そこで、このような、重複する意味を持つ効果語は、同じ形態素を含む場合が多い(「機械特性」と「機械的性質」においては、「機械」)ことを利用し、語を構成する形態素を素性としたクラスタリングを行い、まとめ上げを行う。

6. 手法の評価と考察

6.1. 効果語の抽出に関する評価実験と考察

実験対象となる特許文書は、Nonakaらの手法[Nonaka 2009]を用いて抽出したリサイクルしやすい樹脂に関する特許 1,530 件である。精度については、表1に示す。以下に示すように、パターン1の抽出例としては、「～に使用すること(ができる)」など用途を拡大する効果を持つ「用途拡大系」が多くの部分を占めた。一方、パターン2やパターン3の例としては、「クッション性」など性質を示す効果語が抽出された。全体としてきわめて良好な精度で抽出ができた。

表 1. 精度

パターン1の精度	80.2%
パターン2の精度	94.4%
パターン3の精度	92.4%
全体の精度	85.0%

以下に抽出例を示す。

パターン1の抽出例:「使用する」、「利用する」、「用いる」、「リサイクルする」

パターン2の抽出例:「クッション性」、「強度低下」、「通気性」、「柔軟性」、「精度の低下」

パターン3の抽出例:「機械強度」、「耐衝撃性」、「強度」、「リサイクル性」、「吸着能力」

6.2. 効果語の意味的統合に関する評価実験および考察

今回は効果語の意味的統合に関する基礎的検討として、予めクラスタ数が分かっていることを前提に、形態素を素性としたクラスタリングを行い、手法の妥当性を評価する。実験対象となる特許文書は、Nonakaら[Nonaka 2009]の手法を用いて抽出したリサイクルしやすい樹脂に関する特許 1,530 件である。この特許文書から効果語抽出手法を使用して抽出できた効果語のうち 89 件(ノイズ含む)を対象にまとめ上げを行った。素性は、2個以上の効果語に出現する形態素とし(重み付けは特に行わなかった)、クラスタリングを適用した。クラスタリングは、cluto(<http://www-users.cs.umn.edu/~karypis/cluto/>)を使用して行い、手法としては、RB 法、類似度の評価関数には \cos 尺度を使用した。また、クラスタ数は、7 個(事前の調査による意味上の分類、無関係なもの同士もひとつの分類として扱う)とした。その結果、正解率は、 $83/89=0.93$ となった。

エラーは、

(1) ”ノイズ語である「製品」が「製品外観」をはじめとする外観に関するクラスタに入っていた”、

(2) ”同じく「樹脂品」や「樹脂」、「樹脂物」の樹脂に関するノイズ語が樹脂という意味のクラスタが出来た”、

(3) ”逆に、「用途」や「分野」は、同じような意味を持つ(用途を広げる、分野を広げるなど利用の裾野を広げる効果を持つ)が無関係なものを意味するクラスタに属していた(ただし、共通形態素がないため、この結果は致し方ない。)”、

といったものがあった。このように、効果語を抽出する段階のノイズが悪影響を及ぼしたり、手法の限界によりまとめ上げができていないものがあったが、全体的には高い精度でまとめ上げを行えることが分かった。

7. まとめ

本研究では、効果に相当する語(効果語と呼ぶ)の抽出に焦点をあて、自動的かつ網羅的に効果表現を抽出できる酒井らの手法[酒井 2009]に基づき、その表現中から、ノイズを除去し、直接的に効果を表す語(以下、効果語と言う)を抽出することで、効果-技術型パターンマップの軸のうち効果に相当する語を抽出し、さらに、意味的に重複する語を統合する手法の基礎的検討を行った。

手法の精度について評価したところ、効果語の抽出、および、意味的統合について、いずれも、高い精度を示した。

謝辞

本研究の一部は、文部科学省グローバル COE プログラム、日本学術振興会科研費基盤 C、電気通信普及財団、及び人工知能財団の支援に基づいて行われた。

参考文献

- [Uchida 2004] Hideyuki Uchida, Atsushi Mano and Takashi Yukawa, “Patent Map Generation Using Concept-Based Vector Space Model”, Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, 2004.
- [石川 2004] 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.
- [西山 2008] 西山莉紗, 竹内広宣, 渡辺日出雄, 那須川哲哉, 武田浩一, ”技術文書マイニングのための特長表現抽出”, 第 22 回人工知能学会全国大会, pp.3K-2, 2008.
- [酒井 2009] 酒井浩之, 野中尋史, 増山繁, ”特許明細書からの技術課題情報の抽出”, 人工知能学会論文誌, Vol.24, No.6, pp.531-540, 2009.
- [Nonaka 2009] Hirofumi Nonaka, Hiroyuki Sakai, Shigeru Masuyama, Masazumi Ao, and Yutaka Hada, ”Development of an automatic eco-technology”, patent search method using text-mining”, Journal of Echotechnology Research, Vol.8, pp74-85, 2009.