Vol.2009-NL-192 No.14
2009/7/23

特許文書からのブートストラップ手法を用いた課題・効果表現対の抽出

坂地泰紀[†1] 野中尋史[†1]
酒井浩之[†1] 増山繁[†1]

特許文書から直接的なユーザの便益に相当する表現と，技術上の解決課題を示す表現を自動的に抽出するアルゴリズム「*Cross-Bootstrapping*」を提案する．抽出した直接的なユーザの便益に相当する表現と，技術上の解決課題を示す表現はパテントマップを生成するために役立つ．本手法は，二つの手がかりと統計情報を用いて，ブートストラップ的に表現対を抽出する．また，辞書や人手により作成したパターンを用いず，自動的に表現を抽出することができる．最後に本手法の評価実験を行い，パテントマップを自動生成するために，十分な性能を達成したことを確認した．

Extracting Solution-Effect Expressions from Patent Documents via a Bootstrapping Method

Hiroki Sakaji ,[†1] Hirofumi Nonaka ,[†1]
Hiroyuki Sakai [†1] and Shigeru Masuyama[†1]

We propose a method *Cross-Bootstrapping* for extracting solution-effect expressions from patent documents automatically. Solution-effect expressions are useful for generating patent maps. Our method extracts expressions using two clues and statistical information via a bootstrapping method. Furthermore, the advantage of our method is to extract expressions without dictionaries and patterns given manually. Finally, we experimented our method. As a result, our method achieved sufficient performance for generating patent maps.

†1 豊橋技術科学大学
Toyohashi University of Technology

1. はじめに

パテントマップは，特許の出願動向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，出願目的（特許発明の技術開発がなされた目的を示す）や技術課題（特許発明を使用することにより解決される課題を示す），解決手段（特許発明の構成要素を示す）を，それぞれ軸として，特許出願動向を可視化したもの（以下，上記パテントマップと略す）は，下記のように，特に有益である．技術課題や出願目的は利用者のニーズを示し，解決手段は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・弱み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．一方，特許庁における特許の審査が，技術課題と解決手段，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，自社特許と同じ技術課題・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．さらに，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，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．そのため，上記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．

しかしながら，現状では，技術課題等を特許文書から自動的に抽出する技術が実用化されていないため，前記パテントマップの作成は特許庁等が専門家を利用し，手作業で行っている．そこで，本研究では，このうちの技術課題の抽出に着目し研究を行う．ただし，**技術課題**は，直接的なユーザの便益に相当する表現（「**効果**」と定義する.）と，直接的な便益を実現するために行った技術上の課題解決方法（「**技術上の解決課題**」と定義する.）に分かれる．例を示すと,「本発明によれば，粘着性物質の付着を防止することができ，メンテナンスを最少限に済すことができる.」という技術課題文があった場合,「粘着性物質の付着を防止する」という表現が「技術上の解決課題」であり,「メンテナンスを最少限に済すことができる.」という表現が効果に相当する．また，技術上の解決課題は，解決手段を含んでいる．詳細な解析を行うためには，二つを分けて解析する必要がある．そこで，本研究では，技術上の解決課題を示す**課題表現**と効果を示す**効果表現**の自動抽出を行うアルゴリズム

「*Cross-Bootstrapping*」を提案する．

2. 課題表現と効果表現

効果表現と課題表現が文章中にどのように現れるかを調査する．2000 年に出願された全ての特許明細書 358,085 件の中から「発明の効果」に該当する文，1,228,893 文を抽出し，その中から無作為に選んだ 100 文を調査に用いる．その結果を表 1 に示す．

表 1 100 文中の効果表現と課題表現の出現の仕方

課題と効果の出現場所	出現回数
2 文にまたがって	3
1 文中に出現	65
両方共出現しない、もしくは、どちらか片方しか出現しない	32

表 1 より，課題表現と効果表現のほとんどが同じ文内に出現することが分かる．また，2 文にまたがって出現するものは，その数が少ないので無視できると考え，本研究では 1 文内に出現する課題表現と効果表現の抽出を目指す．

次に，課題・効果表現が具体的にどのように出現しているかを調べた．課題・効果表現を含む文の例を以下に示す．

移載装置は、旋回により収容部材を搬送することができるため、移載装置を小型化することができ、洗浄システムの据付けに必要な設置スペースを少なくすることができる。

この文では,「移載装置を小型化する」が課題表現,「洗浄システムの据付けに必要な設置スペースを少なくすることができる。」が効果表現となる．図 1 に上記の例の課題表現と効果表現を示す．

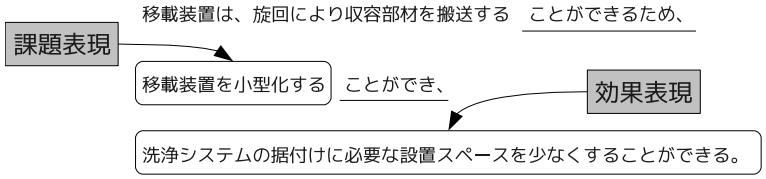

図 1 課題表現と効果表現

図 1 で，下線が引かれた表現「ことができるため、」と「ことができ、」は論理的接続を表す接続指標であると考えられる．また，論理的接続を表す接続指標で区切られた表現「移載装置は、旋回により収容部材を搬送する」「移載装置を小型化する」「洗浄システムの据付けに必要な設置スペースを少なくすることができる。」は連なって，意味を構成していると考えられる．本研究では，これらの論理的接続を表す接続指標に区切られている表現の最後尾のものを効果表現，その直前の表現を課題表現と定義する．パテントマップを生成するうえで，文末に出現する表現が最も重要な情報であり，かつ，その前に出現する表現は，それを補足説明する表現にすぎないため，本研究では上記のような定義とした．

3. 手がかり表現

本研究では,「ことにより、」などの課題・効果表現を抽出するうえで手がかりとなる表現 (以下，**手がかり表現**と定義する) を用いて課題・効果表現を抽出する．

移載装置は、旋回により収容部材を搬送することができるため、移載装置を小型化することができ、洗浄システムの据付けに必要な設置スペースを少なくすることができる。

例えば，以上の文では,「ことができ、」と「ことができる。」が手がかり表現となる．課題表現と効果表現の間に現れ，課題表現の直後に出現する「ことができ、」などの手がかり表現を**課題手がかり表現**と定義する．文末に現れ，効果表現の末尾を構成する「ことができる。」などの手がかり表現を**効果手がかり表現**と定義する．前節の「論理構造を表す接続指標」が課題手がかり表現にあたる．また，課題手がかり表現の末尾には必ず読点を含み，効果手がかり表現の末尾には必ず句点を含むものとする．

3.1 効果手がかり表現の種類

手がかり表現にどのようなものがあるか調べたところ，効果手がかり表現に 2 種類存在した．効果手がかり表現には,「ことができる。」や「ことが優れている。」などがあり，それらを**こと型**と定義した．また,「を図れる。」や「を減少できる。」などがあり，それらを**動詞型**と定義した.「こと型」・「動詞型」効果手がかり表現の種類数を調査した．調査は節 2 で抽出した 65 個の課題・効果表現対を対象とした．結果を表 2 に示す．

表 2 65 個の効果手がかり表現の場合分け

効果手がかり表現の種類	出現回数
こと型	42
動詞型	23

表 2 より，こと型効果手がかり表現の方が，動詞型効果手がかり表現より多いことが分か

る．この結果を用いて，後述する動詞型効果手がかり表現の抽出を行う．

4. 提案手法

課題・効果表現対を抽出するために，課題手がかり表現と効果手がかり表現をブートストラップ的に自動的に獲得する手法を提案する．手がかり表現を二つ使うことと，目的の表現を獲得するために，特徴的な動詞や名詞を経由させることと，抽出の様子が複雑に交差していることから，アルゴリズム *Cross-Bootstrapping* と呼ぶこととする．以下にその手続きを示し，図 2 にその概要を示す．

[*Cross-Bootstrapping*]

Step 0 $S \leftarrow \emptyset$, $E \leftarrow \emptyset$

Step 1 初期課題手がかり表現をいくつか選び課題手がかり表現集合 S の要素とする．また，初期効果手がかり表現をいくつか選び効果手がかり表現集合 E の要素とする．

Step 2 課題手がかり表現集合 S と効果手がかり表現集合 E(こと型効果手がかり表現のみ) を用いて，課題動詞を獲得する．課題手がかり表現集合 S と効果手がかり表現集合 E を用いて，効果動詞と効果名詞を獲得する．

Step 3 獲得した課題動詞と効果手がかり表現集合 E(こと型効果手がかり表現のみ) を用いて，新たな課題手がかり表現を獲得する．獲得した効果動詞・名詞と課題手がかり表現集合 S を用いて，新たな効果手がかり表現を獲得する．

Step 4 新たに獲得した課題・効果手がかり表現をそれぞれ S と E に追加する．

Step 5 Step 2 から 4 を予め定められた回数繰り返す．

Step 6 課題手がかり表現集合 S と効果手がかり表現集合 E を用いて課題・効果表現対を抽出する． □

課題動詞，効果動詞・名詞の定義については後の節で説明する．また，Step 2 と Step 3 に関しては，節 4.1 から節 4.6 で説明する．

Cross-Bootstrapping の特徴としては，課題・効果の二つの手がかりと関連性の高い語を獲得し，それを用いて手がかりを獲得することにある．このようにすることで，二つの手がかりを用いて，互いの手がかりを獲得することができ，また，関連性の高い語を経由させることで様々な手がかりを獲得することができる．その結果，獲得される手がかりは，課題・効果の互いに関連性の強いものとなり，精度の向上も見込める．Pantel et al.[4] や Thelen et al.[5] などの既存のブートストラップ手法では，いずれも手がかりは一つしか用いておらず，また，名詞を対象としているものがほとんどである．手がかりを二つ用いたブートストラップ手法は，我々の知る限りでは，本稿で掲載する *Cross-Bootstrapping* だけである．

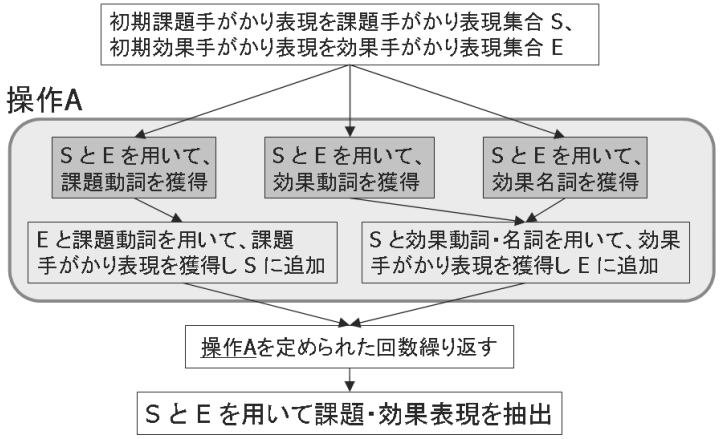

図 2 Cross-Bootstrapping の概要

4.1 課題動詞の獲得

課題手がかり表現を獲得するために，課題動詞を獲得する．**課題動詞**とは，課題手がかり表現の直前に出現し，課題手がかり表現と共起しやすい動詞句と定義する．

移載装置は、旋回により収容部材を搬送することができるため、移載装置を小型化することができ、洗浄システムの据付けに必要な設置スペースを少なくすることができる。

上記の例では，課題手がかり表現「ことができ、」の直前に出現する「を小型化する」が課題動詞となる．

課題動詞は，課題・効果手がかり表現が存在する文から，課題手がかり表現の前に出現する文字を格助詞が出現するところまでで切り取ったものである．具体的には，図 3 に示すようなパターンを作り，これとパターンマッチングして獲得する．

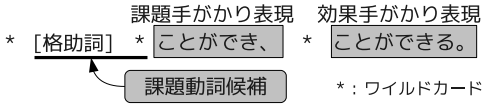

図 3 課題動詞の獲得

Vol.2009-NL-192 No.14
2009/7/23

図 3 のようなパターンとパターンマッチングして獲得した表現を課題動詞候補とする．ただし，課題動詞候補の中には課題手がかり表現を獲得することには不適切なものも含まれているため，選別を行う．

様々な課題・効果手がかり表現対と共起する課題動詞は，課題手がかり表現を獲得する上で有用であるという仮定に基づき，スコアに課題・効果手がかり表現対と課題動詞の共起確率によるエントロピーを用いる．5 回以上抽出された課題動詞候補に対して，以下の式 (1) を用いてスコアを計算する．スコアは 0 から 1 の値を取るように正規化している．

$$Score(s_v) = \frac{H(s_v)}{\log_2 |S||E|} \tag{1}$$

$$H(s_v) = -\sum_{e_c \in E} \sum_{s_c \in S} P_{s_v}(e_c, s_c) \log_2 P_{s_v}(e_c, s_c) \tag{2}$$

$$P_{s_v}(e_c, s_c) = \frac{f_{s_v}(e_c, s_c)}{N(s_v)} \tag{3}$$

ただし，

S: 課題手がかり表現集合

E: 効果手がかり表現集合

$P_{s_v}(e_c, s_c)$: 課題動詞候補 s_v が効果手がかり表現 e_c と課題手がかり表現 s_c と共起する確率

$f_{s_v}(e_c, s_c)$: 課題動詞候補 s_v と効果手がかり表現 e_c と課題手がかり表現 s_c の共起数

$N(s_v)$: 課題動詞候補 s_v の獲得数

スコアが閾値 α 以上のものを課題動詞として獲得する．

4.2 課題手がかり表現の獲得

課題動詞と効果手がかり表現を用いて課題手がかり表現を獲得する．課題手がかり表現は課題動詞と効果手がかり表現が存在する文から，課題動詞の後に出現する文字を読点が出現するところまでで切り取ったものである．具体的には，図 4 に示すようなパターンを作り，これとパターンマッチングして獲得する．

図 4 のようなパターンとパターンマッチングして獲得した表現を課題手がかり表現候補とする．課題手がかり表現候補の中には不適切なものも含まれているため，選別を行う必要がある．ここでも，様々な課題動詞と効果手がかり表現に共起する課題手がかり表現候補は，適切であるという仮定に基づき，スコアに課題動詞・効果手がかり表現と課題手がかり表現の共起確率によるエントロピーを用いる．5 回以上抽出された課題手がかり表現候補に

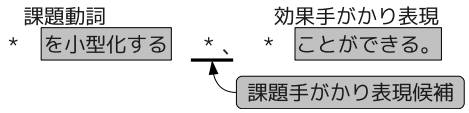

図 4 課題手がかり表現の獲得

対して，以下の式 (4) を用いてスコアを計算する．スコアは 0 から 1 の値を取るように正規化している．

$$Score(s_c) = \frac{H(s_c)}{\log_2 |S_v||E|} \tag{4}$$

$$H(s_c) = -\sum_{e_c \in E} \sum_{s_v \in S_v} P_{s_c}(e_c, s_v) \log_2 P_{s_c}(e_c, s_v) \tag{5}$$

$$P_{s_c}(e_c, s_v) = \frac{f_{s_c}(e_c, s_v)}{N(s_c)} \tag{6}$$

ただし，

S_v: 効果動詞集合

$P_{s_c}(e_c, s_v)$: 課題手がかり表現候補 s_c が課題動詞 s_v と効果手がかり表現 e_c と共起する確率

$f_{s_c}(e_c, s_v)$: 課題手がかり表現候補 s_c と課題動詞 s_v と効果手がかり表現 e_c の共起数

$N(s_c)$: 課題手がかり表現候補 s_c の獲得数

スコアが閾値 α 以上のものを課題手がかり表現として獲得する．

ただし，以下の語を含むものは除く．

ともに 共に とき 時 場合 際 なく 無く ない こと、
だけ と、 一方、 など、 前に、

課題・効果表現間の関係が別の意味に移ってしまうのを防ぐために，これらの語が含まれている課題手がかり表現を除いている．例えば,「とき」や「場合」,「際」などが含まれていると，課題・効果表現間の意味が,「ある指定した条件の時に可能なこと」に変わってしまう．

また,「でき、」などの直前に出現する形態素が名詞である課題手がかり表現は本手法では獲得できない．そこで，課題手がかり表現「でき、」は人手で追加する．

4.3 効果動詞の獲得

こと型効果手がかり表現を獲得するために，効果動詞を獲得する．**効果動詞**とは，効果手がかり表現の直前に出現し，効果手がかり表現と共起しやすい動詞句と定義する．下記の例では，

効果手がかり表現「ことができる。」の直前に出現する「を少なくする」が効果動詞となる．

移載装置は、旋回により収容部材を搬送することができるため、移載装置を小型化することができ、洗浄システムの据付けに必要な設置スペースを少なくすることができる。

効果動詞は，課題・効果手がかり表現が存在する文から，効果手がかり表現の前に出現する文字を格助詞が出現するところまでで切り取ったものである．具体的には，図 5 に示すようなパターンを作り，これとパターンマッチングして獲得する．

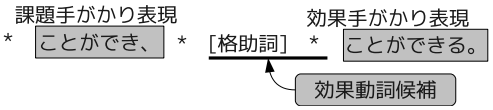

図 5 効果動詞の獲得

図 5 のようなパターンとパターンマッチングして獲得した表現を効果動詞候補とする．効果動詞候補の中には効果手がかり表現を獲得するには不適切なものも含まれているため，選別を行う必要がある．

様々な課題・効果手がかり表現対と共起する効果動詞は，効果手がかり表現を獲得する上で有用であるという仮定に基づき，スコアに課題・効果手がかり表現と効果動詞の共起確率によるエントロピーを用いる．式については，課題動詞と同様であるため，割愛する．スコアが閾値 α 以上のものを効果動詞として獲得する．

4.4 効果動詞を用いた「こと型」効果手がかり表現の獲得

効果動詞と課題手がかり表現を用いて，こと型効果手がかり表現を獲得する．効果手がかり表現は効果動詞と課題手がかり表現が存在する文から，効果動詞の後に出現する文字を句点が出現するところまでで切り取ったものである．具体的には，図 6 に示すようなパターンを作り，これとパターンマッチングして獲得する．

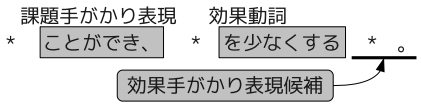

図 6 効果動詞を用いた効果手がかり表現の獲得

図 6 のようなパターンとパターンマッチングして獲得した表現を効果手がかり表現候補とする．効果手がかり表現候補の中には不適切なものも含まれているため，選別を行う必要がある．ここでも，様々な効果動詞と課題手がかり表現に共起する効果手がかり表現候補は，適切であるという仮定に基づき，スコアに効果動詞・課題手がかり表現と効果手がかり表現の共起確率によるエントロピーを用いる．式については，課題手がかり表現と同様であるため，割愛する．スコアが閾値 α 以上のものを効果手がかり表現として獲得する．

4.5 効果名詞の獲得

動詞型効果手がかり表現を獲得するために，効果名詞を獲得する．**効果名詞**とは，効果手がかり表現の直前に出現し，効果手がかり表現と共起しやすい名詞と定義する．下記の例では，効果手がかり表現「を図れる。」の直前に出現する「向上」が効果名詞となる．

光量が最小となる再帰性反射体からの反射光が、光学的開口面に略垂直に入射されるようにしたので、光量が最小となる反射光を効率良く受光でき、検出精度の向上を図れる。

効果名詞は，課題・効果手がかり表現が存在する文から，効果手がかり表現の前に出現する名詞である．具体的には，図 7 に示すようなパターンを作り，これを用いて獲得する．

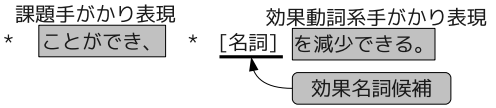

図 7 効果名詞の獲得

図 7 のようなパターンとパターンマッチングして獲得した表現を効果名詞候補とする．5 回以上獲得された効果名詞候補を効果名詞として獲得する．

4.6 効果名詞を用いた「動詞型」効果手がかり表現の獲得

効果名詞と課題手がかり表現を用いて，動詞型効果手がかり表現を獲得する．効果手がかり表現は効果名詞と課題手がかり表現が存在する文から，効果名詞の後に出現する文字を句点が出現するところまでで切り取ったものである．ただし，文字列は助詞から開始していることと，動詞を一つだけ含むことを条件とする．具体的には，図 8 に示すようなパターンを作り，これとパターンマッチングして獲得する．

図 8 のようなパターンとパターンマッチングして獲得した表現を効果手がかり表現候補とする．効果手がかり表現候補の中には不適切なものも含まれているため，選別を行う必要がある．ここでも，様々な効果名詞と課題手がかり表現に共起する効果手がかり表現候補

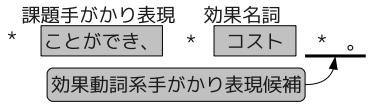

図 8 効果名詞を用いた効果手がかり表現の獲得

は，適切であるという仮定に基づき，スコアに効果名詞・課題手がかり表現と効果手がかり表現の共起確率によるエントロピーを用いる．5 回以上抽出された効果手がかり表現候補に対して，以下の式 (7) を用いてスコアを計算する．スコアは 0 から 1 の値を取るように正規化してある．

$$Score(e_{vc}) = \frac{H(e_{vc})}{max_{H(e_{vc})}} \tag{7}$$

$$H(e_{vc}) = -\sum_{e_n \in E_n} \sum_{s_c \in S} P_{e_{vc}}(e_n, s_c) \log_2 P_{e_{vc}}(e_n, s_c) \tag{8}$$

$$P_{e_{vc}}(e_n, s_c) = \frac{f_{e_{vc}}(e_n, s_c)}{N(e_{vc})} \tag{9}$$

ただし，

E_n: 効果名詞集合

$max_{H(e_{vc})}$: すべてのエントロピー $H(e_{vc})$ の中で最大のもの

$P_{e_{vc}}(e_n, s_c)$: 効果手がかり表現候補 e_{vc} が効果名詞 e_n と課題手がかり表現 s_c と共起する確率

$f_{e_{vc}}(e_n, s_c)$: 効果手がかり表現候補 e_{vc} と効果名詞 e_n と課題手がかり表現 s_c の共起数

$N(e_{vc})$: 効果手がかり表現候補 e_{vc} の獲得数

スコアが閾値 α 以上のものを効果手がかり表現として獲得する．

ここで，動詞型効果手がかり表現を獲得する場合においてのみ，スコアの正規化方法を候補の中でエントロピーの値が最大のもので割るようにしている．これは,「こと」とは違い,「名詞 + 動詞」は決まった形で出てくることが多いので，これに対応するため，上記のような計算を行った．例えば,「コスト + を削減する」は出現するが,「コスト + を小型化する」は出現しない．

4.7 課題・効果表現の抽出

獲得した課題・効果手がかり表現を用いて，課題・効果表現の対を抽出する．課題手がかり表現と効果手がかり表現を同時に含む文に対して，以下の手続きを実行して，課題・効果表現対を抽出する．

Step 1 複数の課題手がかり表現を含む場合，適切な手がかり表現を決定する．最も文末近くに出現し，かつ，最長の文字列になる表現を適切な課題手がかり表現として採用する．効果手がかり表現においても，文字列が最長のものを適切な効果手がかり表現として採用する.(図 9 を参照．)

Step 2 適切な課題手がかり表現から，文頭に向かって文節を結合していき，適切な手がかり表現より後続の文節に係る文節までを課題表現候補として抽出する．

Step 3 課題表現候補中に課題手がかり表現が含まれるなら，課題手がかり表現と，それより前の文字列を削除し，残った文字列を課題表現として抽出する．

Step 4 適切な課題手がかり表現と効果手がかり表現の間の文字列に，効果手がかり表現を結合した文字列を効果表現として抽出する． □

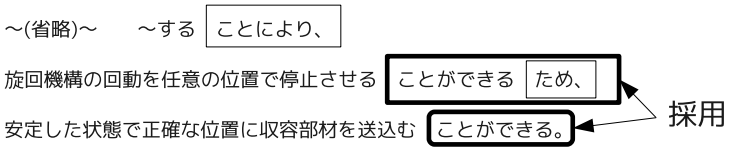

図 9 適切な手がかり表現の採用

5. 評 価 実 験

本手法の性能を評価するために，評価実験を行った．2000 年に出願された全ての特許明細書 358,085 件の中から「発明の効果」に該当する文，1,228,893 文を抽出し，それを評価実験に用いた．正解データとして，上記の 1,228,893 文から無作為に 100 文選び，人手でタグを付与したものを用いた．上記の「発明の効果」100 文にタグを付与したところ，60 個の課題・効果表現対が存在した．形態素解析器には Mecab[⋆1] を用い，係り受け解析器には Cabocha[1] を用いた．初期手がかり表現には, 表 3 を用いた.

課題・効果表現を機械的に正解データと完全に一致しているかを判定すると，意味は同じであるが，長さが少し異なるだけで不正解としてしまう．そこで，抽出した課題・効果表現が正しいかどうかは人手で判断した．適合率 (P)，再現率 (R)，F 値 (F-Measure) の定義を

⋆1 http://mecab.sourceforge.net/

表 3 初期手がかり表現のリスト

課題手がかり表現	ことにより、　ことができ、
「こと型」効果手がかり表現	ことができる。　ことができるようになった。
「動詞型」効果手がかり表現	を減少できる。　を図れる。

以下に示す.

$$P = \frac{|A|}{|Q|}, \quad R = \frac{|A|}{|T|}, \quad F\text{-}Measure = \frac{2PR}{P+R} \tag{10}$$

ただし,

A: 正解データから本手法によって抽出した課題，もしくは，効果表現のうち，正解であった課題，もしくは，効果表現を要素とする集合

Q: 正解データから本手法によって抽出した課題，もしくは，効果表現を要素とする集合

T: 正解データに含まれる人手で抽出した 60 個の課題・効果対を要素とする集合

また，初期手がかり表現だけを使用して，節 4.7 の手続きを用いるものベースラインとする.

5.1 評価結果

表 5 に，ベースラインの結果を示す．表 6〜9 に，本手法の結果を示す．また，図 4 に，各閾値 α とループ回数時の手がかり表現数を示す.

表 5 から 9 より，閾値 $\alpha = 0.5$ でループ回数 3 回のとき，両方の F 値が 0．90 と最も高かった．逆に，ベースラインにおける両方の F 値が 0.19 と最も低かった．これにより，自動的に手がかり表現を獲得する本手法の有用性を示すことができたと考えられる．また，F 値 0．90 という高い値を達成することができた．これは，テキストマイニングにおいて，非常に優れた結果であり，また，パテントマップの自動生成のためには，十分な性能であると考えられる.

評価結果において，本手法は閾値が低いほどよい結果になっている傾向が見られた．これは，以下の 3 つの要因に起因すると考えられる．一つ目は，スコア付けに用いたエントロピーがうまく働いて，不適切な表現を除去できたことである．二つ目は，二つの手がかり表現を用いたことである．本手法は二つの手がかり表現を用いるため，どちらかの手がかり表現に不適切なものがあった場合，4.7 節の手続きにおいて，文にマッチせず課題・効果表現を抽出しない．そのため，不適切な手がかり表現が含まれていても，その影響を抑えることができる．三つ目は，4.7 節において，適切な手がかり表現が獲得できていない場合，他の手がかり表現を適切な手がかり表現であると判断してしまうことである．以上の 3 つの要因により，本手法の結果がもたらされたと考えている.

表 4 使用した手がかり表現の数一覧

α	ループ数	課題手がかり表現	効果手がかり表現		総数
			こと型	動詞型	
ベースライン		2	2	2	6
0.7	1	2	2	18	22
0.7	2	2	2	20	24
0.7	3	2	2	20	24
0.7	4	2	2	20	24
0.7	5	2	2	20	24
0.6	1	4	3	24	31
0.6	2	7	7	29	43
0.6	3	37	28	31	96
0.6	4	37	28	50	115
0.6	5	37	28	81	146
0.5	1	19	44	29	92
0.5	2	22	164	95	281
0.5	3	30	218	185	433
0.5	4	33	245	207	485
0.5	5	37	245	211	493
0.4	1	101	74	31	206

表 5 ベースラインの評価結果

	課題表現			効果表現			両方		
抽出数	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
9	0.78	0.11	0.19	0.78	0.11	0.19	0.78	0.11	0.19

表 6 閾値 α が 0.7 であるときの評価結果

		課題表現			効果表現			両方		
ループ数	抽出数	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
1	14	0.79	0.18	0.30	0.93	0.22	0.35	0.79	0.18	0.30
2	14	0.79	0.18	0.30	0.93	0.22	0.35	0.79	0.18	0.30
3	14	0.79	0.18	0.30	0.93	0.22	0.35	0.79	0.18	0.30
4	14	0.79	0.18	0.30	0.93	0.22	0.35	0.79	0.18	0.30
5	14	0.79	0.18	0.30	0.93	0.22	0.35	0.79	0.18	0.30

6. 関連研究

情報検索システム評価用テストコレクション構築プロジェクト (NTCIR) において，NTCIR-

表 7 閾値 α が 0.6 であるときの評価結果

		課題表現			効果表現			両方		
ループ数	抽出数	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
1	16	0.94	0.25	0.39	0.94	0.25	0.39	0.94	0.25	0.39
2	20	0.95	0.32	0.48	0.95	0.32	0.48	0.95	0.32	0.48
3	36	0.97	0.58	0.73	0.97	0.58	0.73	0.97	0.58	0.73
4	38	0.97	0.62	0.76	0.97	0.62	0.76	0.97	0.62	0.76
5	40	0.98	0.65	0.78	0.98	0.65	0.78	0.98	0.65	0.78

表 8 閾値 α が 0.5 であるときの評価結果

		課題表現			効果表現			両方		
ループ数	抽出数	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
1	34	0.94	0.53	0.68	0.97	0.55	0.70	0.94	0.53	0.68
2	45	0.96	0.72	0.82	0.98	0.73	0.84	0.96	0.72	0.82
3	**51**	**0.98**	**0.83**	**0.90**	**1.00**	**0.85**	**0.92**	**0.98**	**0.83**	**0.90**
4	52	0.94	0.82	0.88	0.96	0.83	0.89	0.94	0.82	0.88
5	53	0.94	0.83	0.88	0.96	0.85	0.90	0.94	0.83	0.88

表 9 閾値 α が 0.4 であるときの評価結果

		課題表現			効果表現			両方		
ループ数	抽出数	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
1	39	0.92	0.60	0.73	0.97	0.63	0.77	0.92	0.60	0.73

4 より特許マイニングのタスクが設定され，NTCIR-4 においては Patent Map Generation タスクが設定された[2]．また，NTCIR-7 においては，特許マイニングタスクとして，日本語または英語論文抄録を特許分類体系のひとつである「国際特許分類 (IPC)」に自動分類するタスクを設定している[3]．このタスクで開発された技術は，将来的には，同一 IPC が付与された特許と論文を，例えば「要素技術」と「効果」という 2 つの観点で再分類し，「要素技術」と「効果」を軸にもつ「技術動向マップ」を作成することに利用できる．

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[6]．本手法においても，課題・効果表現の抽出に有用な手がかり表現を使用することで課題・効果表現の抽出を行う．しかしながら，石川らの手法では「ことにより」を使用していない文から因果関係を抽出することができないが，本手法では自動的に獲得した手がかり表現 433 個があるため，ほとんどの場合に対応することができる．

7. ま と め

特許文書から直接的なユーザの便益に相当する表現と，技術上の解決課題を示す表現を自動的に抽出する手法を提案した．抽出した直接的なユーザの便益に相当する表現と，技術上の解決課題を示す表現はパテントマップを生成するために役立つ．本手法は，二つの手がかりと統計情報を用いて，ブートストラップ的に表現対を抽出する．最後に本手法の評価実験を行い，パテントマップを自動生成するために，十分な性能であることを確認した．

8. 今後の課題

パテントマップを自動生成するために，表現をカテゴリごとに分類する必要がある．また，アルゴリズム「*Cross-Bootstrapping*」が他の表現抽出に用いることができないか検討する．

謝辞 本研究は文部科学省グローバル COE プログラム「インテリジェントセンシングのフロンティア」の支援を受けた．

参 考 文 献

1) 工藤 拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌， Vol.43, No.6, pp.1834–1842 (2002).
2) Fujii, A., Iwayama, M. and Kando, N.: Test Collections for Patent-to-Patent Retrieval and Patent Map Generation in NTCIR-4 Workshop, *in Working Notes of NTICR-4* (2004).
3) Nanba, H., Fujii, A., Iwayama, M. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-7 Workshop, *in Proceeding of NTCIR-7 Workshop Meeting*, pp.325–332 (2008).
4) Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *in Proceedings of the 21st Internatinal Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp.113–120 (2006).
5) Thelen, M. and Riloff, E.: A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts, *in Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp.214–221 (2002).
6) 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).