Cross-Bootstrapping：特許文書からの課題・効果表現対の自動抽出手法

坂地　泰紀[†a)]　野中　尋史[†b)]　酒井　浩之[†c)]　増山　繁[†d)]

Cross-Bootstrapping: An Automatic Extraction Method of Solution-Effect Expressions from Patent Documents

Hiroki SAKAJI[†a)], Hirofumi NONAKA[†b)], Hiroyuki SAKAI[†c)], and Shigeru MASUYAMA[†d)]

あらまし　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．特許出願件数は年間 40 万件にものぼり，1 文書当りの文章量も膨大であるため，出願動向調査に有用なパテントマップ（特許出願動向を可視化したもの）を手作業で作成するには多大な時間とコストを要するため，その作成に役立つ情報を自動的に抽出する技術が求められている．そこで，本研究ではパテントマップの作成に役立つ「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．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，F 値 0.89 と高い性能を達成したことを確認した．

キーワード　パテントマイニング，情報抽出，テキストマイニング，ブートストラップ手法

1. まえがき

日本国内における特許出願は年間約 40 万件にも達している．特許出願される技術は，出願人である企業等にとって非常に重要なものである場合が多く，その出願動向を調査することは，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大きく寄与する．パテントマップは，特許出願動向を可視化したものであり，出願傾向を容易に把握することができ，上記調査の際に重宝されている．しかし，年間の出願件数も多大であり，かつ，特許 1 件当りの文書量も膨大であるため，人手によるパテントマップの作成にはばく大な作業量が掛かるという問題がある．そのため，パテントマップを自動的に作成する技術の開発が求められている．

ここで，特許庁作成のパテントマップ(注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）や発明の効果（特許発明を使用することによりもたらされる効果を示す），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，後述のように，特に有益である．発明の効果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，双方について，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，発明の効果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，自社特許と同じ発明の効果・解決手段をも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．更に，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

† 豊橋技術科学大学，豊橋市
Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan
a) E-mail: sakaji@smlab.tutkie.tut.ac.jp
b) E-mail: nonaka@smlab.tutkie.tut.ac.jp
c) E-mail: sakai@smlab.tutkie.tut.ac.jp
d) E-mail: masuyama@tutkie.tut.ac.jp

(注1)：http://www.jpo.go.jp/shiryou/s_sonota/tokumap.htm

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，重点分野にもかかわらず，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，発明の効果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，発明の効果と解決手段の対の抽出に着目し，研究を行う．

ここで，本論文では，発明の効果は，直接的なユーザの便益に相当するため効果と呼び，解決手段は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であるため技術上の解決課題と呼ぶ．また，効果と技術上の解決課題は，本研究では複数の単語で構成される表現として獲得する．なぜなら，例えば，「コスト」という名詞だけを獲得したとしても，後続する文字列が「が上昇する」と「を抑える」では意味が異なってくるので，複数の単語で構成される表現として獲得する必要があるためである．効果と技術上の解決課題の例を示すと，「単一の熱可塑性材料から形成されているため、リサイクルが可能である。」という文があった場合，「単一の熱可塑性材料から形成されている」という表現が「技術上の解決課題」であり，「リサイクルが可能である。」という表現が効果に相当する．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をともに自動抽出するアルゴリズム「*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 年に出願されたすべての特許明細書 358,085 件の中から「発明の効果」に該当する文，1,228,893 文を機械的に抽出し，その中から無作為に選んだ 200 文を調査に用いる．その結果を表 1 に示す．本研究では「発明の効果」の項目を対象に研究を行う．特許明細書には，「課題」または「発明が解決しようとする課題」という必須項目があり，1 文程度でその特許の効果が簡潔に説明されている．しかし，本研究は技術上の解決課題と効果の両方を抽出することを目的としている．したがって，技術上の解決課題も記述されている「発明の効果」から課題・効果表現を抽出する方が妥当である．そのため，本研究では「発明の効果」タグに該当する文集合から課題・効果表現の抽出を行う．

表 1　効果表現と課題表現の位置の分類
Table 1　A classification of positions of solution-effect expressions in patent documents.

課題と効果の出現場所	出現回数
2 文にまたがって出現	10
1 文中に出現	129
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	61

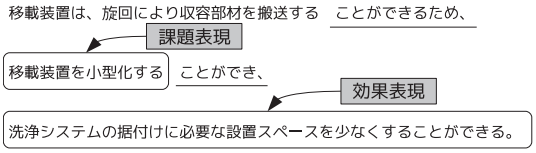

図 1　課題表現と効果表現
Fig. 1　Solution and effect expressions.

表 1 より，課題表現と効果表現の約 65%が 1 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 文内に同時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 では，下線が引かれた表現「ことができるため、」と「ことができ、」は因果関係 [1] を表す接続表現である．本論文では，これらの接続表現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である可能性が高いと仮定し，これに対応する手法を考案する．このようにした理由はパテントマップを生成する上で，「発明の効果」に該当する文では，文末に出現する表現と，その直前に出現する表現が最も重要な情報であり，また，表 1 の調査結果の 1 文中に出現した課題・効果表現対 129 個すべてにおいて，文末には効果表現が，その直前には課題表現

が出現していたためである．

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上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，上記の文では，「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．前節の接続表現が課題手がかり表現に該当する．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，効果手がかり表現は 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「こと」で始まっているものがあり，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。」など助詞で始まり，次に動詞がきているものがあり，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 **2.** で抽出した 129 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 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．

表 2 効果手がかり表現の分類
Table 2 A classification of effect expressions.

効果手がかり表現の種類	出現回数
こと型	93
動詞型	36

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．本手法では，異なる二つの手がかり表現と，目的の表現を獲得するために，特徴的な動詞や名詞を用いている．そして，特徴的な動詞や名詞を利用して，課題・効果手がかり表現を獲得している様子が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．

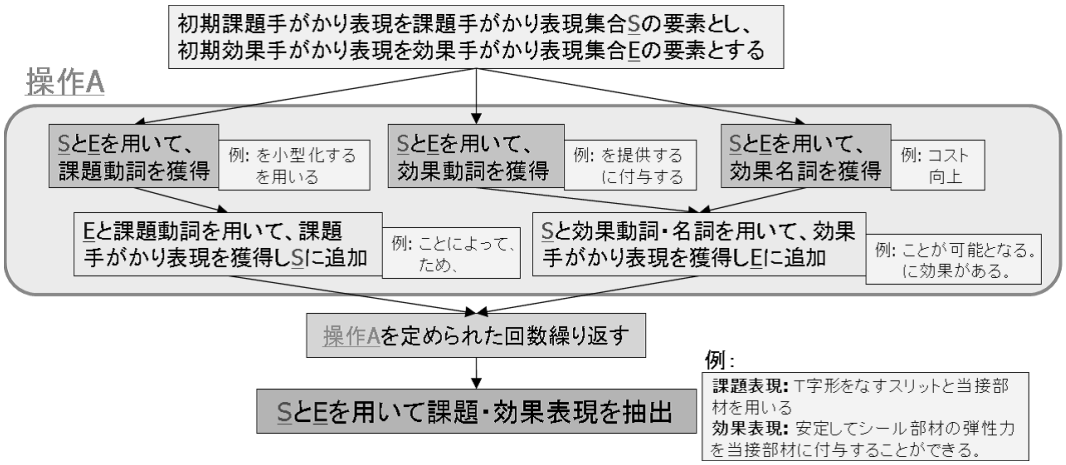

図 2 Cross-Bootstrapping の概要
Fig. 2 An outline of Cross-Bootstrapping.

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し，S に追加する．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し，E に追加する．
Step 4　Step 2 から 3 をあらかじめ定められた回数繰り返す（図 2 において，Step 2 と 3 が操作 A に当たる）．
Step 5　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，**4.1** から **4.6** で説明する．

Step 2 と 3 において，課題動詞を獲得するときと，課題手がかり表現を獲得するときに，こと型効果手がかり表現だけを用いているのは，以下の理由からである．予備実験において，課題動詞や課題手がかり表現を獲得する際に，こと型効果手がかり表現だけでなく，動詞型効果手がかり表現も用いた場合，うまく獲得することができなかった．これは，課題手がかり表現の数に対して，効果手がかり表現の数が圧倒的に多いことで，Step 2 や 3 で用いているエントロピーの値が小さくなり，課題動詞や課題手がかり表現を獲得できなくなってしまったためである．また，逆に，動詞型効果手がかり表現だけを用いた場合も，課題手がかり表現に対して動詞型手がかり表現の数が少ないために，多種の課題動詞を獲得できず，結果的に少数の課題手がかり表現しか獲得することができなかった．以上のことから，こと型効果手がかり表現だけを用いることにした．

Cross-Bootstrapping の特徴は，課題・効果の二つの手がかり表現と関連性の高い語を獲得し，それを用いて手がかり表現を獲得することである．本手法のブートストラップ部分では，抽出目的である課題・効果表現ではなく，それを獲得するための課題・効果手がかり表現を獲得している．そのため，この部分だけを見れば，獲得すべき表現は互いに関連するが異なる 2 種類の手がかり表現であると考えることができる．また，本手法では，課題手がかり表現を獲得する際に，効果手がかり表現をパターンに用い，効果手がかり表現を獲得する際には，課題手がかり表現を用いている．更に，関連性の高い語（課題・効果表現を構成する一部分である可能性が高い語）を介在させることにより，Step 2 と 3 を通して見たときに，課題手がかり表現を獲得する際の入力に課題手がかり表現を加えることができる．以上より，表現を獲得するパターンを精密化でき，獲得される手がかり表現は課題・効果の互いに関連性の強いものとなり，精度の向上が見込める．

なお，課題表現と効果表現を別々に抽出するような手法も考えられる．このような場合には，1 文内に抽出した課題表現と効果表現が存在すれば，それらが対であると判断するであろう．しかしながら，例えば，「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。」という文の場合，課題表現として「旋回により収容部材を搬送することができる」，効果表現として「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。」を抽出していたとしたら，それらが対であると判断してしまう．確かに，それらの間には関係はあるが，パテントマップを作成する上では，効果表現に，より関連性が高いものが重要であるため，課題表現として「移載装置を小型化する」の方が妥当であると考えられる．ほかにも，1 文内に含まれていても表現間には関係がないものも考えられ，そのようなものを除外するため，本手法のように課題・効果表現対として獲得する必要がある．

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ることにより獲得することができる．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．

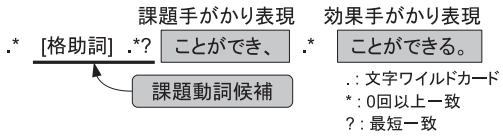

図 3 課題動詞の獲得
Fig. 3 An extraction of solution verbs.

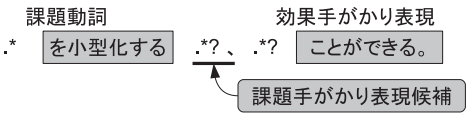

図 4 課題手がかり表現の獲得
Fig. 4 An extraction of solution clue expressions.

パターンは課題手がかり表現と，効果手がかり表現の集合から一つずつ要素を取り出して，総あたりで組を作成する．以下の表現抽出においても同様である．また，課題動詞候補には課題手がかり表現の直前に出現する格助詞を含める．動詞と格助詞を合わせることによって，課題動詞の意味が決まってくるため，格助詞を含めて獲得する．例えば，「を調節する」と「が調節する」では，目的の物を調節するという意味と，主体が調節するという意味になり，格助詞の種類によって意味が異なる．我々は，例えば，前者の目的の物を調節するという意味の「を調節する」の方が「が調整する」より，課題表現を構成する末尾として出現する可能性が高いのであれば，これを獲得したいというねらいから，このような獲得方法にした．効果動詞に関しても同様の理由で格助詞を含めている．

課題動詞候補を獲得するパターンに課題手がかり表現だけではなく，効果手がかり表現も用いることによって，図 3 の課題動詞候補は課題表現の動詞句部分となる可能性が高くなる．そのことにより，次に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ときに，課題手がかり表現を獲得しやすくなると考え，課題・効果手がかり表現の両方をパターンに用いることにした．効果名詞，効果動詞の獲得においても同様である．

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，課題動詞の選別を行う．

ここで，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，適切な課題動詞を獲得するためのスコアに課題・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．なお，スコアは 0 から 1 の値をと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

スコアがしきい値 α 以上の課題動詞候補を課題動詞として獲得する．

特許文書の中には筆者の記述誤りが散見される．それを除去するために 5 回以上抽出された課題動詞だけをスコア付けの対象とする制限をかけている．事前に調査したところ，この制限を設けることでほとんどの記述誤りを取り除くことができたため，本論文では 5 回以上抽出できた課題動詞候補をスコア付けの対象としている．以下の課題・効果表現を除く表現抽出においても同様の理由で，5 回以上という制限を設けている．

また，事前調査により 1 文字の課題動詞候補は不適切であることを確認したため，課題動詞候補として獲得してくる文字列は，2 文字以上の文字列に限定している．以下の表現抽出においても同様の理由で 2 文字以上のものを獲得するという制限を設けている．

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．図 4 に示す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．

課題動詞の獲得と同様の理由で，エントロピーを用

いて課題手がかり表現の選別を行う．式については，課題動詞の獲得に用いた式と同様であるため，割愛する．スコアがしきい値 α 以上の課題手がかり表現候補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、

これらの語が課題手がかり表現に含まれると，表現間の意味が因果関係以外となり，適切に課題・効果表現を獲得できない．そこで，それを防ぐために，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．これらの語は，課題表現と効果表現の調査に用いた 200 文から人手で獲得したものである．上記の語が課題手がかり表現に含まれた場合において，表現間の意味ごとに分類したものを表 3 に示す．表 3 より，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．

本手法では，ストップワードが含まれているものを除くという方法をとっているため，以下のような語も除外できる．

ことなく、 際にも、 できるとともに、 際に、のみでなく、 必要がなく、 ことがなく、

また，課題・効果表現の中には，ストップワードを含むものが存在する．例えば，「取付孔を利用し、取付クランプの一対の取付片部をボディに取付固定すれば、エアバッグのガス流入部をインフレーターに連結させるとともに、インフレーターをボディに取付固定することができる。」などは，効果表現中に並列を意味する「とともに、」を含む．このようなストップワードを含む課題・効果表現を **2.** で調査に用いた 200 文から抽出し，その数を調査した．結果を表 4 に示す．

200 文中には，課題・効果表現対が 129 あり，そのうち 21 の表現対は並列や否定などを表す語を含んでいた．これらの語を適切に除外しておかないと，後述する *Expression Extraction* において，正しく課題・効果表現を獲得できない．

また，「でき、」などの直前に出現する形態素が名詞や助詞である課題手がかり表現は本手法では獲得できない．なぜならば，効果動詞とあるように，本研究で用いる課題手がかり表現の直前に出現する形態素の品詞は動詞なので，直前に出現する形態素の品詞が名詞や助詞である「でき、」は課題動詞の直後には出現せず，獲得することができない．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。

図 5 に示す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．

ここでも，課題動詞や課題手がかり表現の獲得と同様の理由で，エントロピーも用いて効果動詞の選別を行う．式についても，課題動詞と同様であるため，割愛する．スコアがしきい値 α 以上の効果動詞候補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

表 3 ストップワードの分類
Table 3 A classification of stop words.

意味	ストップワード
指定した条件の時	とき 時 場合 際 前に、
並列	ともに 共に と、
否定	なく 無く ない
限定	だけ
逆接	一方、
その他	こと、 など、

表 4 ストップワードを含む課題・効果表現の数
Table 4 The number of solution-effect expressions including stop words.

意味	表現対数
指定した条件の時	4
並列	9
否定	8

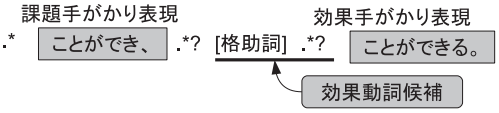

図 5 効果動詞の獲得
Fig. 5 An extraction of effect verbs.

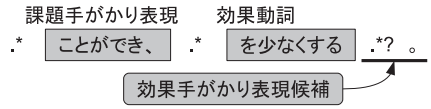

図 6　効果動詞を用いた効果手がかり表現の獲得
Fig. 6　An extraction of effect clue expressions using effect verbs.

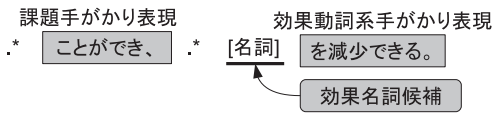

図 7　効果名詞の獲得
Fig. 7　An extraction of effect nouns.

手がかり表現を獲得する．図 6 に示す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．

ここでも，課題動詞の獲得と同様の理由で，エントロピーを用いて効果手がかり表現の選別を行う．式についても，課題動詞の獲得や課題手がかり表現の獲得と同様であるため，割愛する．スコアがしきい値 α 以上の効果手がかり表現候補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。

図 7 に示す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ることにより獲得することができる．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．図 8 に示す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．

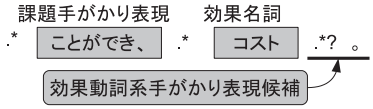

図 8　効果名詞を用いた効果手がかり表現の獲得
Fig. 8　An extraction of effect clue expressions using effect nouns.

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，効果手がかり表現の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は適切であるという仮定に基づき，適切な効果手がかり表現を獲得するためのスコアに効果名詞・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 (4) を用いてスコアを計算する．スコアは 0 から 1 の値をとるように正規化している．

$$Score(e_{vc}) = \frac{H(e_{vc})}{max_{H(e_{vc})}} \tag{4}$$

$$H(e_{vc}) = -\sum_{e_n \in E_n} \sum_{s_c \in S} P_{e_{vc}}(e_n, s_c) \log_2 P_{e_{vc}}(e_n, s_c) \tag{5}$$

$$P_{e_{vc}}(e_n, s_c) = \frac{f(e_{vc}, e_n, s_c)}{N(e_{vc})} \tag{6}$$

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

スコアがしきい値 α 以上の効果手がかり表現候補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全効果名詞候補中の最大のエントロピー値で割るようにしている．これは，以下の理由による．「こと」とは異なり，「名詞 + 動詞」は決まった形で出現することが多い．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．そのため，効果名詞と動詞型効果手がかり表現の共起関係は偏っている．動詞型効果手がかり表現を獲得するために用いるエントロピーの取り得る最大の値 $\log_2 |S||E_n|$ で $H(e_{vc})$ を正規化してしまうと，この偏りの影響でスコアが全体的に小さくなり，動詞型効果手がかり表現を獲得できなくなってしまう．これに対応するため，上記のようなスコアの計算を行った．

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 *Expression Extraction* を実行する．

[*Expression Extraction*]

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，最長の文字列であ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 においては，最も文末近くに出現する課題手がかり表現が「でき、」と「ことができ、」の二つ存在するが，最長の文字列という条件から，「でき、」より文字列の長い「ことができ、」を適切な課題手がかり表現として採用している）．

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，特定の文節（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．

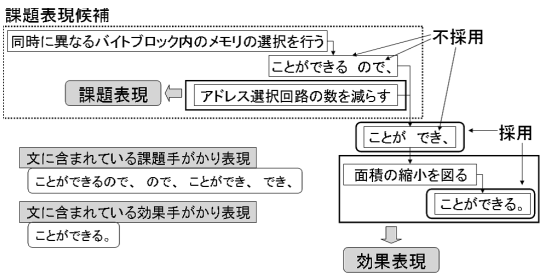

図 9　*Expression Extraction* の実行例
Fig. 9　An execution example of *Expression Extraction*.

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 に *Expression Extraction* の実行例を示す．まず，Step 1 でこの文において適切な手がかり表現を採用する．図 9 では，適切な課題手がかり表現として「ことができ、」，適切な効果手がかり表現として「ことができる。」を採用している．次に Step 2 で課題表現候補を抽出する．図 9 では，「同時に異なるバイトブロック内のメモリの選択を行うことができるので、アドレス選択回路の数を減らす」を課題表現候補として抽出している．次に，Step 3 で課題表現を抽出する．図 9 では，課題表現候補の中に課題手がかり表現「ことができるので、」と「ので、」が含まれているため，それより前の文字を削除し，残った「アドレス選択回路の数を減らす」を課題表現として抽出している．最後に，Step 4 で効果表現を抽出する．図 9 では，適切な課題手がかり表現「ことができ、」の適切な効果手がかり表現「ことができる。」を含めた後続の文字列「面積の縮小を図ることができる。」を効果表現として抽出している．

5. 評価実験

5.1 実験方法

本手法の性能を評価するために，評価実験を行った．2000 年に出願されたすべての特許明細書 358,085 件の中から「発明の効果」に該当する文，1,229,893 文を抽出し，調査に用いた 200 文と正解データに用いた 400 文を除いた 1,229,293 文を実験に用いた．正解データとして，上記の 1,228,893 文から調査に用いた 200 文とは異なる 400 文を無作為に選び，人手で課題・効果表現を示すタグを付与したものを用いた．上記の「発明の効果」400 文にタグを付与したところ，222 個の課題・効果表現対が存在した．形態素解析器としては Mecab(注2) を用い，係り受け解析器としては Cabocha [2] を用いた．初期手がかり表現としては，表 5 を用いた．

課題・効果表現に対して機械的に正解データと完全に一致しているか否かを判定すると，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否かは人手で判断した．以下の判断の基準を示すガイドライン

(注2)：http://mecab.sourceforge.net/

表 5　初期手がかり表現のリスト
Table 5　A list of initial clue expressions.

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。

に基づき正解か否かを判定した．

- 課題・効果表現の直前に出現する接続詞が含まれていても正解とする．

例えば，人手でタグをつけた課題表現「メモリ容量の超過状態を防止できる」に接続詞「また、」が先頭に加わった表現「また、メモリ容量の超過状態を防止できる」が本手法により課題表現として抽出されていたら正解とする．

- 「本発明によれば」や「本発明は」という表現が，課題表現の先頭に加えられていても正解とする．

例えば，人手でタグをつけた課題表現「データベースを分割格納する記憶領域の割り付けを複数種類のレンジで管理する」に「本発明によれば、」が先頭に加わった表現「本発明によれば、データベースを分割格納する記憶領域の割り付けを複数種類のレンジで管理する」が本手法により課題表現として抽出されていたら正解とする．

- 「. . . ことなく、」などの否定表現や，「. . . の際に、」などの条件や場合を表す表現が課題表現の先頭に加えられていても，正解とする．

例えば，人手でタグをつけた課題表現「走査幅のばらつきを抑える」の先頭に「光学部品の組付けの際の位置精度を厳しくすることなく、」が加わった表現「光学部品の組付けの際の位置精度を厳しくすることなく、走査幅のばらつきを抑える」が本手法により課題表現として抽出されていたら正解とする．

- 課題表現の末尾に「構成である」などの重要な情報をもっていない表現が含まれていても正解とする．

例えば，人手でタグをつけた課題表現「その短時間信号を用いて音声認識をする」の末尾に「構成である」が含まれた表現「その短時間信号を用いて音声認識をする構成である」が本手法により課題表現として抽出されていたら正解とする．

適合率 (P)，再現率 (R)，F 値 (F-Measure) の定義を以下に示す．

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

ただし，

A：正解データから本手法によって抽出した課題，若しくは，効果表現のうち，正解であった課題，若しくは，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，若しくは，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 222 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，**4.7** の手続きを用いるものをベースラインとする．

5.2　Pantel らの手法との比較

Pantel らの提案している Espresso は，パタン(注3) p を用いて二つの名詞対 $i = x, y$ を獲得する手法である [3]．我々は，Pantel の手法を課題・効果表現の抽出に適用するために拡張し（以下，拡張 Espresso），本手法との比較を行った．課題動詞と課題手がかり表現，効果動詞とこと型効果手がかり表現，効果名詞と動詞型効果手がかり表現をそれぞれ Espresso におけるインスタンスとパタンとおく．そのため，課題動詞や課題手がかり表現の獲得方法は，課題動詞の獲得であれば，本手法で用いているパターンを効果手がかり表現を含めないものにし，それを各々に適用した．また，本手法のストップワードと，課題・効果表現抽出部分「*Expression Extraction*」を拡張 Espresso に適用した．

課題動詞，効果動詞，効果名詞それぞれの候補の信頼度を式 (7) で計算し，課題手がかり表現，こと型効果手がかり表現，動詞型手がかり表現それぞれの候補の信頼度を式 (8) で計算する．

$$r_l(i) = \frac{\sum_{p \in P} \frac{pmi(i,p)}{max_{pmi}} r_\pi(p)}{|P|} \tag{7}$$

$$r_\pi(p) = \frac{\sum_{i \in I} \frac{pmi(i,p)}{max_{pmi}} r_l(i)}{|I|} \tag{8}$$

ここで，P は各手がかり表現候補，それぞれの集合，I は課題動詞，効果動詞，効果名詞，それぞれの集合である．また，次の式 (9) で $pmi(i,p)$ を計算する．

$$pmi(i,p) = \log \frac{P(i,p)}{P(i)P(p)} \times \frac{C_{ip}}{C_{ip}+1} \times \frac{min(\sum_{k=1}^{|P|} C_{ik},\ \sum_{j=i}^{|I|} C_{jp})}{min(\sum_{k=1}^{|P|} C_{ik},\ \sum_{j=i}^{|I|} C_{jp}) + 1} \tag{9}$$

(注3)：Pantel の手法における Pattern と本手法におけるパターンを区別するために，Pantel のものをパタンとする．

ここで，C_{ip} は，p と i が共起する頻度である．

まず，本手法と同様に六つの手がかり表現を初期手がかり表現として信頼度 r_l を求め，その上位 N 個を課題動詞，若しくは，効果動詞，効果名詞として獲得する．次に，手がかり表現候補の信頼度 r_π を求め，その上位 N 個を各手がかり表現として獲得する．これを定められた回数（本実験では 10），繰返し，課題・効果手がかり表現を獲得する．ここで，N の値をそれぞれ 100 とした．

また，N の値を課題動詞と課題手がかり表現では 30 とし，効果動詞とこと型効果手がかり表現では 218，効果名詞と動詞型効果手がかり表現では 185 としたものも評価した．この値は，本手法で比較的 F 値が高かった，しきい値 0.5 で反復回数 3 のときに獲得された課題手がかり表現，こと型効果手がかり表現，動詞型効果手がかり表現，それぞれの数である．そのため，本手法により学習した N の値を用いていると考えられるので，比較としては必ずしも妥当ではないが，考察のため評価を行った．

5.3 評価結果と考察

表 8 に，各しきい値 α と反復回数ごとの手がかり表現数と，本手法とベースラインの評価結果を示す．表 7 に，しきい値 $\alpha = 0.5$ で反復回数 7 のとき，本手法によって獲得された手がかり表現の一部を示す．表 8 において，ベースラインの手がかり表現数が 7 になっているのは，**4.2** で述べているように課題手がかり表現「でき､」を追加しているためである．

また，しきい値 0.7 において，反復回数 2 以降，新たに手がかり表現を獲得できなかったため，評価結果が変化しなかった．そのため，それ以降は省略している．しきい値 0.6 と 0.4 においても，反復回数 6 以降では，新たに手がかり表現が獲得できなかったため，省略している．しきい値 0.3 の場合に関しては，約 1 か月間 Xeon2 Ghz，Memory32 G のマシンで動作させたが，処理が終わらず，反復回数 2 までの結果となった．

表 6 に，拡張 Espresso の評価結果（課題・効果表現対を評価）を示す．$N = 100$ の場合は，反復回数 2 のとき最も高い F 値となったため，そのときの評価値を記載している．$N = 30,\ 218,\ 185$ の場合は，反復回数 3 のとき最も高い F 値となったため，そのときの評価値を記載している．なお，比較のため，表 6 には本手法において最も F 値の良いもの（しきい値 0.5，反復回数 7）を併記している．

表 6　拡張 Espresso の評価結果
Table 6　Results of Extended Espresso.

	適合率	再現率	F 値
本手法	0.98	0.81	0.89
拡張 Espresso ($N = 100$)	0.89	0.72	0.80
拡張 Espresso ($N = 30,\ 218,\ 185$)	0.97	0.79	0.87

表 8 より，しきい値 $\alpha = 0.5$ で反復回数 7 のとき，課題・効果表現対の F 値が 0.89 と最も高かった．逆に，ベースラインにおける課題・効果表現対の F 値が 0.25 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 0.89 という高い値を達成することができた．

評価結果において，しきい値 0.5 の F 値（課題・効果表現対）の場合，反復回数を増やすごとに上昇していき，反復回数 7 のときを頂点として，以降の反復回数では下がっていく傾向が見られた．また，このときの適合率に着目すると，反復回数 7 以前の適合率の変化はほとんどなく，反復回数 8 以降で下がっていた．これは，反復回数 8 以降で獲得した手がかり表現に不適切なものが含まれ始めたからである．例えば，反復回数 8 以降では，課題手がかり表現として「ように､」を獲得していたため，文「以上説明したことからも明らかなように、本発明ではつぎのような利点がある。」から，誤って課題表現として「以上説明したことからも明らかな」，効果表現として「本発明ではつぎのような利点がある。」を獲得してしまっていた．

また，しきい値を下げ，反復回数を多くしたとしても，適合率が 0.80 以上と高い値になっている．これは，二つの異なる手がかり表現を用いたことによる．本手法は二つの異なる手がかり表現を用いているため，いずれかの手がかり表現に不適切なものがあった場合，**4.7** の手続きにおいて，手がかり表現対が文にマッチせず課題・効果表現を抽出しない．例えば，次のような表現「状態で､」が課題手がかり表現として獲得されていた場合，獲得しているどの効果手がかり表現とパターンを作成しても，正解データ中ではマッチしなかった．そのため，不適切な手がかり表現が含まれていたとしても，その影響を抑えることができたと考えられる．

表 8 より，しきい値 0.7，反復回数 1 と 2 の場合と，しきい値 0.6，反復回数 1 と 2 の場合において，適合率がしきい値 0.6，反復回数 3 以降に比べ低くなっている．これは，課題・効果表現を抽出するときに誤って

表 7　獲得した手がかり表現の例
Table 7　An example of clue expressions.

課題手がかり表現	ことで、 ことが可能であり、 ようにしたため、 構成としたので、 ことができ、 ことができるので、 ようにしているので、 ことができるため、 ために、 ことができるから、 ことができる、 ため、 ことによって、 から、 ものであるから、 ことが可能となり、
「こと型」効果手がかり表現	ことができる効果を有している。 という効果が得られる。 ことが出来るのである。 ことが可能となる。 ことができることである。 ことができるものとなった。 ことができるといった優れた効果を奏する。 ことができる等の特長を有する。 ことが可能となるという効果を奏する。 ことができるという有利な効果が得られる。
「動詞型」効果手がかり表現	に寄与する。 にも貢献できる。 を維持できる。 を軽減できる。 に対応できる。 を防止できる。 が抑制できる。 を構成できる。 が実現できる。 が回避できる。 を生じることがない。 が少なくなる。 の効果がある。 を製造できる。 が提供できる。

表 8　本手法の評価結果
Table 8　Evaluation results of our method.

				課題表現			効果表現			課題・効果表現対		
α	反復回数	手がかり表現数	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
ベースライン		7	42	0.79	0.15	0.25	0.86	0.16	0.27	0.79	0.15	0.25
0.7	1	23	57	0.81	0.21	0.33	0.96	0.25	0.39	0.81	0.21	0.33
0.7	2	25	57	0.81	0.21	0.33	0.96	0.25	0.39	0.81	0.21	0.33
0.6	1	32	66	0.86	0.26	0.40	0.88	0.26	0.40	0.86	0.26	0.40
0.6	2	44	74	0.89	0.30	0.45	0.91	0.30	0.45	0.89	0.30	0.45
0.6	3	97	134	0.99	0.59	0.74	0.99	0.59	0.74	0.99	0.59	0.74
0.6	4	116	140	0.99	0.62	0.76	0.99	0.62	0.76	0.99	0.62	0.76
0.6	5	147	147	0.99	0.65	0.79	0.99	0.65	0.79	0.99	0.65	0.79
0.6	6	154	148	0.99	0.66	0.79	0.99	0.66	0.79	0.99	0.66	0.79
0.5	1	93	120	0.97	0.52	0.68	0.97	0.52	0.68	0.97	0.52	0.68
0.5	2	282	160	0.95	0.68	0.80	0.95	0.70	0.80	0.95	0.68	0.80
0.5	3	434	181	0.97	0.79	0.87	0.97	0.79	0.87	0.97	0.79	0.87
0.5	4	486	182	0.97	0.79	0.87	0.97	0.79	0.87	0.97	0.79	0.87
0.5	5	494	183	0.97	0.80	0.87	0.97	0.80	0.87	0.97	0.80	0.87
0.5	6	503	184	0.97	0.81	0.88	0.97	0.81	0.88	0.97	0.81	0.88
0.5	**7**	**513**	**184**	**0.98**	**0.81**	**0.89**	**0.98**	**0.81**	**0.89**	**0.98**	**0.81**	**0.89**
0.5	8	531	195	0.91	0.80	0.85	0.92	0.81	0.86	0.91	0.80	0.85
0.5	9	585	195	0.92	0.81	0.86	0.92	0.81	0.86	0.92	0.81	0.86
0.5	10	703	200	0.88	0.79	0.83	0.88	0.79	0.83	0.88	0.79	0.83
0.4	1	207	154	0.93	0.64	0.76	0.94	0.65	0.77	0.93	0.64	0.76
0.4	2	549	193	0.91	0.79	0.85	0.92	0.80	0.86	0.91	0.79	0.85
0.4	3	776	201	0.91	0.82	0.87	0.92	0.83	0.87	0.91	0.82	0.87
0.4	4	818	204	0.91	0.83	0.87	0.92	0.84	0.88	0.91	0.83	0.87
0.4	5	827	204	0.91	0.83	0.87	0.92	0.84	0.88	0.91	0.83	0.87
0.4	6	828	204	0.91	0.83	0.87	0.92	0.84	0.88	0.91	0.83	0.87
0.3	1	459	166	0.87	0.65	0.74	0.87	0.65	0.75	0.87	0.65	0.74
0.3	2	1436	213	0.84	0.81	0.82	0.85	0.81	0.83	0.84	0.81	0.82

本来適切でない手がかり表現を適切な手がかり表現として採用してしまうためである．**4.7** の手続きにおいて，獲得した手がかり表現の中に適切な手がかり表現が含まれていない場合，獲得している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がある．その場合，適切な課題・効果表現ではない文字列を抽出してしまい，エラーとなる．例えば，十分に手がかり表現を獲得できていなかったしきい値 0.7 の場合，「2 個の油圧モータの斜板角度の最小値を 2 種設定し、該斜板は、油圧ピストンまたは電動アクチュエータ等の駆動機構で駆動され、最小・最大の 2 位置設定でき、操作ハンドルの 4 位置に対応させて 4 種の

容量を設定したので、変速機の操作を容易に行うことができる。」という文からは，課題表現として「油圧ピストンまたは電動アクチュエータ等の駆動機構で駆動され、最小・最大の 2 位置設定でき、」，効果表現として「操作ハンドルの 4 位置に対応させて 4 種の容量を設定したので、変速機の操作を容易に行うことができる。」が獲得された．しかしながら，適切な課題表現は「操作ハンドルの 4 位置に対応させて 4 種の容量を設定した」，適切な効果表現は「変速機の操作を容易に行うことができる。」である．これは，しきい値 0.7 の場合に，「ので、」が獲得されていなかったため，*Expression Extraction* において「でき、」がこの文において適切な課題手がかり表現として選ばれてしまったからである．この現象は十分な数の手がかり表現を獲得できていないときに起こりやすいため，獲得した手がかり表現の数が少ない場合，適合率が低くなってしまった．

表 8 より，適合率に関しては 0.90 以上の高い値を達成することができたが，再現率に関しては最大でも約 0.80 と適合率に比べ低かった．これは，出現数の少ない手がかり表現を獲得できていないことに起因する．

> 支持線部とケーブルコア部とは窓によって大部分の長さが独立しているので、支持線部とケーブルコア部の進行速度に差を持たせる**ことが容易に出来て、**支持線部の長さよりもケーブルコア部の長さの方が長い所要の余長を有するケーブルを容易に製造することが出来る。

例えば，正解データに含まれていた上記の文における適切な課題手がかり表現「ことが容易に出来て、」は，後述のように出現数が少なく，獲得できていないがために，上記文中の課題・効果表現を抽出することができなかった．今回，実験に用いた 2000 年に出願された特許明細書中の「発明の効果」には，「ことが容易に出来て、」が 1 個だけしか含まれていないため，5 回以上出現という制約を満たせず，破棄されてしまった．このような出現回数の少ない手がかり表現が特許明細書中には多数あり，それらを不適切なものを含めないで獲得できる方法を確立することが今後の課題になる．

表 6 より，精度，再現率，F 値のすべてにおいて，本手法が拡張 Espresso ($N = 100$) を上回った．しかしながら，結果に大きな差が出なかったのは，本手法と同様の手がかり表現等の定義，ストップワード，課題･効果表現抽出部分「*Expression Extraction*」を用いたためであると考えている．拡張 Espresso ($N = 30,\ 218,\ 185$) に関しては，本手法と大差のない結果となった．これは，拡張 Espresso ($N = 30,\ 218,\ 185$) の N の値を，本手法により学習しているためであると考えられる．拡張 Espresso ($N = 100$) において，F 値が低かった理由としては，最適な課題手がかり表現，効果手がかり表現の数が異なっているため，すべての N の値を同一にしてしまうとノイズが入ったり，網羅性が乏しかったりするためである．しかしながら，Espresso では，それぞれの N を人手で与える必要があるため，本手法のように一つのしきい値を定めるだけで自動的に適切な数の課題・効果手がかり表現を獲得することができない．

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-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 [4]．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一つである「国際特許分類 (IPC)」に自動分類するタスクを設定している [5]．

NTCIR-4 におけるパテントマップ自動生成の研究として，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 [6]．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，「発明の効果」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，本論文で提案した手法では抽出する課題・効果表現に含まれるキーワードを自動的に取得することで「発明の効果」「解決手段」に関連したキーワードを取得可能である．

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 [1]．本論文で提案した手法においても，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，本論文で提案した手法では自動的に獲得した手がかり表現 513 個があるため，ほとんどの場合に対応することができる．

また，自然言語処理を利用した特許文書中の表現抽出及び解析技術は，盛んに研究されており，例えば，谷川ら [7] や新森ら [8]，安彦 [9]，難波ら [10] の研究がある．谷川ら [7] の研究では，特許文書中から請求項数などのパラメータを素性として抽出し，教師データとして人手による価値評価結果を与え，機械学習を行い，特許文書の価値を導出している．新森ら [8] の研究では，特許文書中の請求項の読解支援のために，言語処理技術を利用し，請求項の分節や請求項間のトリー関係可視化を行っている．一方，安彦 [9] は，特許請求項の権利範囲を格成分と呼ばれる請求項表現中の限定条件により定量的に扱えることを示している．また，難波ら [10] は，「〜等の …」特許文書中の表現に着目し，語の上位関係を自動的に取得し，オントロジーの構築を行っている．しかし，これらの研究は，技術上の課題や効果を抽出するものではないため，本手法のようにパテントマップ作成にすぐに利用できるものではない．

酒井らは，特許明細書から効果表現を自動的に抽出する手法を提案している [11]．ただし，酒井らの手法は効果表現を抽出できるが，課題表現を抽出することができない．それに対して，本手法では効果表現だけでなく，課題表現も抽出でき，パテントマップを作成する上では酒井らの手法より役立つと考えられる．

ブートストラップ手法を用いた研究としては，Pantel ら [3] や Thelen ら [12] の研究がある．Pantel らや Thelen らなど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，獲得する対象を名詞に限定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，本論文で提案する *Cross-Bootstrapping* が初めてである．

Blum らは，少数のラベル付けされたデータと，二つの学習器を用いてラベルの付与されていないデータにラベルを付与する手法である Co-Training を提案している [13]．Co-Training も本手法と同様にブートストラップ的な手法である．また，類似点を明確にするために，本手法の課題・効果手がかり表現を用いて課題動詞を獲得するステップと，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ステップをまとめて一つの課題手がかり表現の抽出器とする．更に，課題・効果手がかり表現を用いて効果動詞，効果名詞を獲得するステップと，それらと課題手がかり表現を用いて効果手がかり表現を獲得するステップをまとめて一つの効果手がかり表現の抽出器とする．この二つの抽出器を学習器と見れば，二つの学習器を用いている Co-Training の構造と類似している．

しかしながら，本手法は Co-Training のように入力データを分類しているのではなく，入力データ（手がかり表現）を用いて新たな手がかり表現を獲得している点が異なる．加えて，本手法ではパターンマッチングを行って表現を獲得する手法であるため，課題動詞や効果動詞，効果名詞を介在させないと，課題・効果手がかり表現を獲得することができない．この点においても，Co-Training とは異なる．

7. むすび

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 *Cross-Bootstrapping* 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．*Cross-Bootstrapping* は異なる二つの表現を用いて，特徴的な動詞や名詞を獲得し，今度はそれらを用いて手がかり表現を獲得する．その結果，課題・効果表現を抽出するために役立つ手がかり表現を多数獲得でき，それらを用いて高精度に課題・効果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，課題・効果表現対の抽出に関して F 値 0.89 と高い値を達成した．

パテントマップを自動生成するために，表現をカテゴリーごとに分類する必要がある．例えば，「装置全体の小型化を図ることができる」と「発明装置を縮小化することができる」は同じ意味をもつ表現であるが，表層的には異なるため，機械的に同一であると判断することができない．パテントマップを自動作成するためには，関連した特許をまとめる必要があり，そのためには関連した特許間の表現をまとめる必要がある．そこで，特許文書に特化した辞書を作成するなどして，表現のまとめ上げを自動的に行うことが今後の課題となる．

また，特許の特徴を調べるために，抽出した課題・効果表現に含まれる特許を特徴づけるキーワードの抽出が考えられる．例えば，「単一の熱可塑性材料から形成されているため、リサイクルが可能である。」という文では，「熱可塑性材料」という複合名詞が，この文を含む特許を特徴づけるキーワードの一つとなっている．このようなキーワードを抽出することにより，特

許の分析，分類が容易となり，また，上記の表現のまとめ上げにも役立つ．キーワードは課題・効果表現に含まれている可能性が高く，本手法で課題・効果表現を抽出した後に，そこから抽出することで容易にキーワードを得ることができると考えられる．

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」，日本学術振興会科研費基盤（C）22500129，若手（B）21700158，基盤（B）20300058，JST シーズ発掘試験，電気通信普及財団，人工知能研究振興財団，及び，栢森情報科学振興財団の援助により行われた．

文　献

[1] 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.

[2] 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，” 情処学論，vol.43, no.6, pp.1834–1842, 2002.

[3] P. Pantel and M. Pennacchiotti, “Espresso: Leveraging generic patterns for automatically harvesting semantic relations,” Proc. 21st Internatinal Conference on Computational Linguistics and the 44th Annual Meeting of the ACL, pp.113–120, 2006.

[4] A. Fujii, M. Iwayama, and N. Kando, “Test collections for patent-to-patent retrieval and patent map generation in ntcir-4 workshop,” Working Notes of NTICR-4, 2004.

[5] H. Nanba, A. Fujii, M. Iwayama, and T. Hashimoto, “Overview of the patent mining task at the ntcir-7 workshop,” Proc. NTCIR-7 Workshop Meeting, 2008.

[6] H. Uchida and A. Mano, “Patent map generation using concept-based vector space model,” Working Notes of NTICR-4, 2004.

[7] 谷川英和，新森昭宏，“言語処理に基づく特許価値評価支援システムと特許読解支援システム，” 日本知財学会第 4 回年次学術研究発表会，pp.444–449, 2006.

[8] 新森昭宏，奥村　学，丸川雄三，岩山　真，“手がかり句を用いた特許請求項の構造解析，” 情処学論，vol.45, no.3, pp.891–905, 2004.

[9] 安彦　元，“格文法を利用した特許請求の範囲の限定度合解析とその戦略的応用，” 日本知財学会第 7 回年次学術研究発表会，2009.

[10] 難波英嗣，奥村　学，新森昭宏，谷川英和，鈴木泰山，“特許データベースからのシソーラスの自動構築，” 言語処理学会第 13 回年次大会，pp.1113–1116, 2007.

[11] 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，” 人工知能誌，vol.24, no.6, pp.531–540, 2009.

[12] M. Thelen and E. Riloff, “A bootstrapping method for learning semantic lexicons using extraction pattern contexts,” Proc. Conference on Empirical Methods in Natural Language Processing, pp.214–221, 2002.

[13] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” COLT: Proc. Workshop on Computational Learning Theory, pp.92–100, 1998.

（平成 21 年 9 月 5 日受付，22 年 1 月 4 日再受付）

坂地　泰紀

2009 豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻了．現在，同大学院工学研究科博士後期課程電子・情報工学専攻在学中．自然言語処理の研究に従事．

野中　尋史

2005 豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻了．2005 三重大学総合情報処理センター技術員．2006 独立行政法人新エネルギー・産業技術総合開発機構フェロー（〜2008）．2007 豊橋技術科学大学知財連携コーディネーター（〜現在に至る）．2008 豊橋技術科学大学大学院工学研究科博士課程電子・情報工学専攻入学（〜現在に至る）．自然言語処理，特に，特許文書や判例文書の解析に関する研究に従事．情報ネットワーク法学会，知財学会各会員．

酒井　浩之

2002 豊橋技術科学大学大学院工学研究科修士課程知識情報工学専攻了．2005 同大学院工学研究科博士後期課程電子・情報工学専攻了．博士（工学）．2005 豊橋技術科学大学知識情報工学系助手．2007 豊橋技術科学大学知識情報工学系助教．自然言語処理，特に，テキストマイニング，テキスト自動要約の研究に従事．言語処理学会，人工知能学会各会員．

増山　繁　（正員）

1977 京大・工・数理工学卒．1982 同大大学院博士後期課程単位取得退学．1983 同修了（工博）．1982 日本学術振興会奨励研究員．1984 京都大学工学部数理工学科助手．1989 豊橋技術科学大学知識情報工学系講師．1990 同助教授．1997 同教授．2005 豊橋技術科学大学インテリジェントセンシングシステムリサーチセンター教授併任．アルゴリズム工学，特に，並列アルゴリズム等，及び，自然言語処理，特に，テキスト自動要約等の研究に従事．言語処理学会，情報処理学会等各会員．