特許文書からの発明に関する特徴的技術とその効果の抽出

原田綾花 [1]　太田貴久 [1]　小林暁雄 [1]　増山 繁 [1]　野中尋史 [2]　酒井浩之 [3]
[1]豊橋技術科学大学　[2]大分工業高等専門学校　[3]成蹊大学
harada@la.cs.tut.ac.jp　kikyu@la.cs.tut.ac.jp　kobayashi@la.cs.tut.ac.jp
masuyama@tut.jp　h-nonaka@oita-ct.ac.jp　h-sakai@st.seikei.ac.jp

1.　はじめに

知財戦略や研究開発戦略立案の際，特許文書を対象とした技術調査を行う必要がある．しかしながら，多数の特許に記述されている発明を正しく把握する作業は手間と経験を要し，調査対象によっては，知財の専門家であっても 2 ヶ月以上必要な場合がある．そこで，本研究では，技術調査に必要な情報を，特許から自動的に抽出することを目指す．技術調査の際に重要な情報は，特許に記載されている，発明を特徴づける技術（以下，技術）と，発明の効果（以下，効果）である．ここで，効果とは，「小型化」や「省電力化」といった，その発明の利用者の便益を表す発明が解決しようとする課題に該当する．本研究では，発明の技術と効果を把握するために，特許から図 1 に示す 4 つの要素を抽出する．

A. 技術の総称
B. 技術構成要素
C. 技術構成要素の機能/作用
D. 発明の効果

図 1．本研究の抽出対象

図 1 において，B と D が，それぞれ発明を特徴づける技術と，発明の効果を表す．また，A．技術の総称とは，発明全体を表す語であり，「音響装置」や「音声再生システム」などがこれにあたる．さらに，多くの場合，技術構成要素が直接発明の効果をもたらすのではなく，技術構成要素の作用によって，発明の効果を達成する．そこで，本研究では，技術構成要素と発明の効果を結ぶ情報として，C．技術構成要素の機能 / 作用も抽出する．図 2 に，消しゴム付きの筆記具に関する特許について，上記に挙げた 4 つの要素をそれぞれ抽出した例を示す．図 2 のように，A により発明の技術分野の概略を同定することができる．さらに，B，C，および，D を組み入れることで，発明がどのような技術によって構成され (B)，その技術がどのような目的で用いられ (C)，最終的に，発明の利用者にどのような便益をもたらすか (D) を知ることができる．ここで,C と D は，発明の利用者の便益にあたるか否かが異なるのみであり，技術構成要素がもたらす現象という意味では同じカテゴリに属する．

A.　筆記具
B.　鉛筆と、消しゴムと、上記鉛筆の端と上記消しゴムの端とを連結する連結具を備える。
C.　鉛筆と消しゴムとが連結されて一緒にあるため、
D.　消す作業が楽になる。

図 2．抽出例

したがって，本研究では両者をまとめて発明の「作用・効果」として扱う．また，本研究では，自然言語処理，および，非高分子有機化合物である添加剤によって特徴づけられる潤滑組成物の 2 分野（特許に振られている国際特許分類コードが，それぞれ，G06F17/27-28，C10M137/02 ）を対象として，調査，研究を行った．

2.　従来手法

特許文書からの情報抽出に関する研究の多くは，発明を特徴づける技術と効果の抽出を目的としている（図 1 の B，C，D ）．例えば，発明の効果に相当する表現 （以下，効果表現 ）を抽出する手法として，西山らの手法（[1]，[2])や，Nanba らの手法 [3] がある．西山らの手法 [1] では，「頑健性」などの技術にとっての好ましい要素や，「ができる」などの発明の効果を抽出するための手がかりとなる表現（以下，手がかり表現 ）を利用することで，効果表現を抽出していた．Nanba ら [3] の手法でも，「軽減」や「効果」などの手がかり表現を用いることで，技術，および，効果のラベル付けを自動的に行っている．しかしながら，これらの手法では，手がかり表現は人手 / 半自動的に抽出しなければならない．また，体言のみからなる効果表現には対応できていない等，網羅性に欠けていた．一方，酒井らの手法 [4] では，発明の効果タグに該当する文集合から，「ができる。」と「が可能である。」といった表現を，手がかり表現の種として与えることで手がかり表現を自動的に得ることができる．しかしながら，この手法では，効果に相当する表現の中でも，「抽出できる」などのように，助詞「が」を用いない表現や，受身の表現となっている文は抽出することができないという問題がある．技術構成要素を抽出する手法としては，鈴木の手法 [5] が挙げられる．鈴木の手法では，技術構成要素の直後に出現する形態素列 （以下，デリミター ）のうち，エントロピーの高いデリミターの前に出現する形態素列を，技術構成要素として抽出する （図3）．しかしながら，この手法においても，まだ十分な精度と再現率が実現されていない．

したがって，本研究では，精度を保ちつつ，これらの技術と効果に関する表現を漏らさずに，すなわち，再現率よく取得する手法について検討を行った．

一対のピンチローラと、これより下流に設けた**エアーブロー装置**とからなることを特徴とする・・・

図 3．技術構成要素とデリミターの例
（下線：デリミター，太字：技術構成要素）

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

3. 提案手法

3. 1. 提案手法の概要

本研究では，請求項（特許請求の範囲）の末尾の文末に，技術の総称が記述されるという特徴を利用して技術の総称を抽出する手法を提案する．また，技術構成要素の抽出については，技術構成要素の末尾の形態素は，特定の形態素であることが多いため，これらを手がかりとして抽出する手法を提案する．発明の作用・効果に関する内容については，主に，特許明細書中の【発明の効果】，【解決手段】，【課題を解決するための手段】の３つのタグが割り当てられた段落に記述されることが多い．ここで，タグとは，図４のように，明細書に記述されている見出しのことをいう．しかしながら，これらのタグに該当する段落には，図４に示す「本発明によって、以下のような効果が得られる。」といった，作用・効果以外の内容が記述されることも多い．提案手法では，これら３つのタグが振られた段落から，作用・効果以外について記述されている内容を省くことで，作用・効果を抽出する．

【解決手段】 ・・・ （省略） ・・・
【発明の効果】本発明によって、以下のような効果が得られる。

図４． 特許明細書中のタグの例

3. 2. 技術の総称の抽出

本手法では，まず，技術の総称の抽出を行う．請求項は，誤解を招かない記載方法にするため，図５のように，構成要素を先に列挙した上で最後に技術の総称を記載することが多い．したがって，技術の総称は，多くが各請求項の末尾の文末から取得できることが分かる（図５）．そこで，請求項の末尾の文末に出現する表現を，技術の総称として抽出する．

・・・原始データとともに保持する蓄積手段を備えたことを特徴と**する**データ蓄積変換システム。

図５．技術の総称の例（下線：技術の総称）

3. 3. 技術構成要素の抽出

鈴木の手法 [5] では，まず，技術構成要素は，表１の条件を両方とも満たす形態素列として，これらを技術構成要素候補として抽出する．そして，これらの中から，特定のデリミターの前に出現する形態素列を技術構成要素として抽出しているが，それだけでは技術構成要素以外の形態素列も多数抽出してしまう．そのため，「であって」や，「において」などの，従来技術を示すデリミターの前に出現する形態素列を，技術構成要素候補から除外するようにしている．しかしながら，これらのデリミターの前には，例えば図6(b)のように，技術構成要素が記述されることもあるため，再現率低下の原因になってしまう．

そこで，デリミターを手がかりとするのではなく，図6(a)のように，技術の総称が請求項の末尾以外の場所に出現する場合，その技術の総称の出現位置以前に出現する形態素列を，技術構成要素候補から除外するようにした．また，技術構成要素の末尾の形態素は，例えば，「手段」や「工程」など，特定の形態素（以下，手がかり表現 ）であることが多い．そこで，本手法では，技術構成要素候補の末尾の形態素が手がかり表現と一致するものを技術構成要素として抽出する．このときに使用する手がかり表現は，人手で定義し与える．

表１．技術構成要素条件

条件１．
技術構成要素を構成する品詞が，表２のいずれかである．
条件２．
技術構成要素の末尾形態素の品詞が，表３のいずれかである．

表２．技術構成要素を構成する品詞

非自立名詞，代名詞以外の名詞，体言接続動詞，
句読点以外の記号，接頭辞または接尾辞

表３．技術構成要素の末尾形態素の品詞

一般名詞，サ変名詞，固有名詞，
接尾辞，アルファベット

(a) [従来技術(existing technology)]を有した**音響装置**において、[技術構成要素(technology terms)]とを有することを特徴とする**音響装置**。
(b) [技術構成要素]において、・・・

図６．技術構成要素の例（太字：技術の総称）

3. 4. 発明の作用・効果の抽出

次に，発明の作用・効果の抽出を行う．請求項中には，通常，発明の作用・効果は記述されないことから，前記３つのタグ中の文集合から，請求項と一致する内容の文 / 句を省くことで，作用・効果以外の内容の文 / 句を除去できると考える（図7）．また，請求項と一致する内容を省いた後の文 / 句集合から，除去されずに残った作用・効果以外の文 / 句を，SVM で分類を行うことにより省く．これにより，最終的に残った文 / 句から作用・効果を抽出する．この手法において，請求項と一致する内容を省く処理は現時点では未実装のため，人手で行った．

特許請求項

・・・(省略)・・・鉛筆と、消しゴムと、連結具で連結されていることを特徴とする筆記具。

発明の効果タグ

請求項１にあるように、鉛筆１と、消しゴム２と、連結具とを連結しているため、**鉛筆と消しゴムとが連結されて、常に一緒にあるため、消す作業が楽になる。**

図７．請求項（上）と発明の効果タグ（下）の例
（下線：特許請求項の内容と一致する部分
太字：発明の作用・効果）

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

3．4．1　発明の作用・効果の抽出の予備実験

そこで，まず，予備実験として，前記 3 つのタグを対象に，作用と効果に関する文 / 句などが出現する種類の数を調査した．その結果，図 8 に示す全 8 種類の文 / 句の項目に分類された．また，これらの項目について，前記 3 つのタグに該当する文集合中における，文字数の割合を調べた（図 11，12）．その結果，作用・効果に関する文 / 節は，解決手段，および，課題を解決するための手段タグに該当する文集合の文字数においては約40%，発明の効果タグにおいては74%と大部分を占めていることが分かった．それ以外の項目として，(a) 接続詞，文頭，文末や，(o) 技術要素に関する項目は，解決手段，および，課題を解決するための手段タグに該当する文集合においては，それぞれ23%，33%，発明の効果タグに該当する文集合においては，それぞれ12%，１%と，(e) 作用・効果ほどではないものの，各タグ中を占める割合が多いことが分かった．

●発明に関する情報を含む文/句
(e) 作用・効果　（例：強度を上げることができる。）
(f) 作用・効果の条件など（例：旅行中などに、）
(p) 従来例　（例：従来の翻訳装置は・・）
(c) ハードウェア構成（例：入力部はマウスを用いる。）
(d) 単語の定義　（例：ここで述語とは、・・・）
(t) その他，技術要素に関する項目
（例：モデルは、音声データに依存して作成する。）
●発明に関する情報を含まない文 / 句
(a) 接続詞、文頭、文末など（例：さらに、）
(o) 参照を示す文　（例：以下で説明する。）

図 8．特許明細書におけるタグ中の文集合における項目

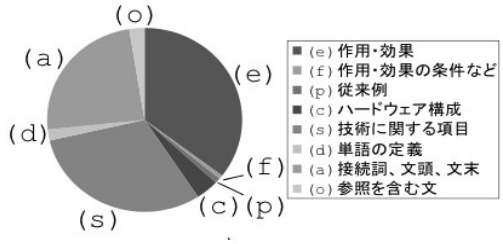

図 9．解決手段，および，課題を解決するための手段タグ中の文集合における項目の文字数の割合

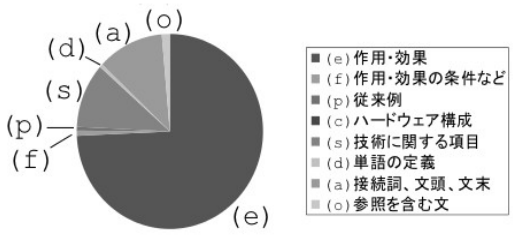

図10．発明の効果タグ中の文集合における項目の文字数の割合

4．　実験および結果

4．1　技術の総称の抽出実験

2002年度に公開された特許 300 件を対象に，本手法によって技術の総称を抽出したところ，精度 93%，再現率 94%を得ることができた．

4．2　技術構成要素の抽出実験

1996年～2006年度に公開された特許のうち，ランダムに抽出した特許 71件（自然言語処理分野 ）を対象に，鈴木の手法を用いて技術構成要素を抽出したところ，精度 43%，再現率 48%であった．これに対し，図 11に示すような手がかり表現（全21件 ）を用意し，技術構成要素候補の末尾が，図 11の手がかり表現であるものを対象として技術構成要素の抽出を行ったところ，精度 68%，再現率 70%と，精度・再現率を大幅に向上させることができた．

一方，自然言語処理以外の分野として， 2002年度に公開された特許 30件（潤滑組成物に関する分野）を対象として，上記と同様に技術構成要素の抽出を行った．その結果，精度 56%，再現率 74%と，こちらも従来手法を上回る結果が得られた．上記の分野では，「剤」や「成分」，「カーボン」など，特に，薬品や化合物などを示す手がかり表現 （全88件 ）が多く，その種類も多様であったため，図 11の手がかり表現と比べ，約 4 倍の手がかり表現が必要であった ．以上のことから，提案手法が分野によらず従来手法を上回る精度，再現率を実現すると予想される．

段階	手段	プログラム
工程	データベース	プロセス

図11．手がかり表現の例（太字：技術の総称）

4．3　発明の作用・効果の抽出実験

表 4 に， SVM-Light(http://svmlight.joachims.org/) による，作用・効果に関する項目と，それ以外の項目との分類精度を調べた結果を示す （正例：作用・効果に関する項目，負例：作用・効果以外の項目 ）．データは， 2002年度に公開された自然言語処理分野の特許文書からランダムに選んだ 100 件を用いた．素性の単位は形態素とし，特徴ベクトルは，(a1)は形態素の有無を 1， 0 で表したもの，(a2)は(a1)に品詞情報を加えたもの，(a3)は形態素の頻度，(a4)は(a3)に品詞情報を加えたものとして，それぞれの特徴ベクトルを用いた際の SVM の分類精度を 10-fold cross validation によって調べた（表4)．(a1)と(a3)との間では，精度，および，再現率にあまり差は見られなかったが，いずれも品詞情報を付与することで，多くの場合，精度，再現率を上げることができた．「解決手段」，「課題を解決するための手段」タグに対しては，(a4)の特徴ベクトルを用いた際が，精度 81%，再現率73%，「発明の効果」タグに対しては， (a1)の特徴ベクトルを用いた際が，精度 95%，再現率 95%を実現することができた．

さらに，表 5 は，各項目の種類に対し，誤って分類された数と割合を示している．これらの結果から，特に精度が良 かった項目としては， (a) 接続詞，文頭，文末，および， (o) 参照を含む文が挙げられ，誤りが多かったものとしては，(s) 技術要素に関する項目が挙げられた．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表４．　SVM による分類精度，および，再現率

対象タグ		(a1)	(a2)	(a3)	(a4)
解決手段，課題を解決するための手段	精度	0.79	0.77	0.79	0.81
	再現率	0.69	0.71	0.71	0.73
発明の効果	精度	0.96	0.95	0.96	0.95
	再現率	0.98	0.96	0.98	0.95

表５．　SVM によって分類を誤った項目の数と割合

項目の種類	各項目の種類に対し，誤って分類された数/総数（割合）
(e) 作用・効果	74/601 (0.12)
(f) 作用・効果の条件など	7/9 (0.78)
(a) 接続詞，文頭，文末	4/525 (0.01)
(s) 技術要素に関する項目	52/140 (0.37)
(o) 参照を含む文	2/46 (0.04)
(c) ハードウェア構成	6/11 (0.55)
(d) 単語の定義	3/8 (0.38)
(p) 従来例	4/6 (0.67)

5.　考察

5．1　技術構成要素の抽出についての考察

本手法で抽出できなかった技術構成要素としては，自然言語処理分野では，「翻訳モデルをトレーニングするステップ」や，「単語列を認定するサブステップ」などのように，技術構成要素に体言接続動詞以外の動詞を含む例が大半を占めていた．これは，本手法の前処理において，技術構成要素を構成する形態素の品詞が限定されており，体言接続動詞以外の動詞は含まれないことが原因である．一方，潤滑組成物に関する分野の特許では，従来技術を技術構成要素として誤って抽出してしまう例が多かった．本手法では，前処理の段階で，「前記」や「該」などが付されている形態素列を，従来技術として技術構成要素候補から除去している．しかし，従来技術であってもそれらの形態素が付随していない特許が当分野では多く，そのために，それらの従来技術を前処理で除外できなかったことが原因である．

5．2　発明の作用・効果の抽出についての考察

表４で示したように，(a) 接続詞，文頭，文末，および，(o) 参照を含む文の，分類精度が良かった理由としては，「そして」や「したがって」などの接続詞や，「次に」や「以下」などの特有の語が多いことから，分類が容易であったものと解釈できる（図13～図 14）．一方，(s) 技術要素に関する項目の，誤りが多かった理由としては，図15に示している例のように，「次に」や「以下」など，項目特有の語が出現しないことから，作用・効果への誤分類が多かったものと解釈できる．その他の，精度が悪かった項目の理由としては，事例の数がまだ少なく，特徴を捉えきれていないためと考えられる．

・また、この発明は、
・といったものがある。

図13．図 9 における項目 (a) の例

・そのための方法を以下に示す。
・上記の通り、本発明には下記のような効果がある。

図14．図 9 における項目 (o) の例

・音響モデルを、ユーザ音声データに依存して作成する。
・連続しているテキストが翻訳単位として切り出される。

図15．図 9 における項目 (s) の例

6.　まとめ

本研究では，技術の総称，技術構成要素，発明の作用・効果を抽出することで，ユーザにとって有益な情報を抽出する手法を提案し，その有効性を検討するための実験を行った．その中でも，技術の総称を高精度に抽出することに成功した．また，技術構成要素の抽出に関しては，自然言語処理と潤滑組成物の 2 つの分野で実験を行った結果，提案手法より，いずれも，従来手法よりも高い再現率，精度が得られた．このことから，本手法は分野に依存することなく従来手法より再現率，精度ともに改善すると予想される．発明の作用・効果を抽出する手法においては，請求項と一致する項目を省いた後の文から，SVM による分類を行うことで，作用・効果を精度・再現率よく抽出できることが分かった．今後は，現段階では実現できていない，請求項と一致している内容を省く処理の実装を行う．また，今後の課題として，技術構成要素の抽出については，技術構成要素に体言接続動詞以外の動詞が含まれる場合の抽出方法の検討が挙げられる．また，従来技術に「前記」や「該」などが付されていない場合の，従来技術の判定方法についての検討が挙げられる．

参考文献

[1] 西山 莉紗，竹内 広宜，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. 3K3-2 (2008)

[2] 西山 莉紗，竹内 広宜，渡辺 日出雄，那須川 哲哉，前田 潤治，倉持 俊之，林口 英治：未来技術動向予測のための技術文書マイニング，第 21 回人工知能学会全国大会予稿集，No. 2H5-3 (2007)

[3] H. Nanba, T. Kondo and T. Takezawa: Hiroshima City University at NTCIR-8 Patent Mining Task, Proceedings of NTCIR Workshop 8 Meeting, 2010.

[4] 酒井 浩之，他，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.

[5] 鈴木 祐輔，特許明細書からの技術構成要素の抽出，平成23年度修士論文，豊橋技術科学大学大学院工学研究科知識情報工学専攻修士課程，January 2010.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.