# 論文と特許からの技術動向マップの自動構築

近藤 友樹　　難波 英嗣　　竹澤 寿幸
広島市立大学大学院 情報科学研究科

## 1.はじめに

近年，大学研究者自身が関連論文だけでなく関連特許について情報を検索することや，特許の出願･分析を行う機会が増えており，2009 年 6 月に政府の知的財産戦略本部が発表した「知的財産権推進計画 2009」においても，大学研究における特許情報の重要性が謳われている．この計画に，大学研究者の利用を想定した特許･論文情報統合検索システムの整備が含まれていることからも，このような傾向は今後さらに強まっていくと思われる．こうした状況を鑑み，著者らは，特許と論文を対象にした技術動向分析支援システムの構築に取り組んでいる．

その一例として，近藤ら[近藤 2007]は日本語の論文表題から，その論文の主題と要素技術の対を抽出し，技術動向マップの自動構築を行っている．しかし，論文表題のみでなく，論文と特許の概要なども対象にすれば，より網羅的で多様な観点からの技術動向分析が可能になると考えられる．本研究では，特定分野の特許と論文概要から，要素技術とその効果を示す表現を自動的に抽出し，技術動向分析を行うシステムについて述べる．

本論文の構成は以下の通りである．次節でシステムの動作例を示す．3 節では，関連研究について述べ，4 節では，本研究での論文および特許概要の構造解析手法を述べる．5 節では，有効性を調べるために行った実験について報告し，6 節で結果を考察する．最後に 7 節で本稿をまとめる．

## 2.技術動向分析システム

著者らは，研究動向を可視化するシステムを構築している．以下に，システムの動作例および仕組みについて説明する．図 1 は，「音声認識」という用語をシステムに入力した時の解析結果を示している．図 1 において，左側に「音声認識」の要素技術が列挙され，各技術の右側にその用語が使われている年が示される．例えば図 1 中にある要素技術「隠れマルコフモデル」の場合，この用語を要素技術に用いた文献が 1999 年に発表されていることを示している．これらは図中で「●」として表示されており，ユーザが●上にカーソルを重ねることで，その文献の書誌情報がポップアップウィンドウ内に表示される．

また，図 1 において要素技術として提示されている用語をユーザがクリックすることで，その要素技術が他にどのような分野で利用されているのかを，年代順に一覧表示することができる．図 2 は，図 1 中の「隠れマルコフモデル」をクリックした結果を示している．学術界では 1990 年代前半に音声認識の分野で使われていた技術が 1990 年代後半に入ると文字認識や画像認識の分野でも利用されていることが一覧表示される．

さらに，各要素技術の効果に関する情報が，各図の右端に表示される．図 1 では，「音声認識」の分野で「概念素」の技術から「無矛盾性を確保」できるという効果が得られることや，「マイクロホンアレー」の技術では「収束が遅くなる」という効果があることが分かる．また，図 2 では，様々な分野においてある要素技術にどのような効果があるのか一覧できる．


図 1.「音声認識」で使われる要素技術と効果の一覧表示


図 2.「隠れマルコフモデル」を要素技術として用いている分野と各分野における効果の一覧表示

## 3.関連研究

### 3.1 表題の構造解析およびその応用

近藤ら[近藤 2008]は，論文の表題に着目して，機械学習により，要素技術と主題の抽出をしている．また，これ以前に日本語論文の表題解析を行っているが，これと同様の手法を用いた英語論文の表題を解析において，「英語論文表題の構造が一般的に日本語のものより複雑である」ことと，「構造解析に用いる手掛かり語が日本語ほど有効に機能しない」ということを指摘している．また，これらの問題を，日本語論文表題の構造解析結果と数種の翻訳方法を使った方法である"翻訳知識"を用いて解決している．

1 節で述べた近藤らの日本語論文表題の構造解析と本節で述べた英語論文表題の構造解析は，論文表題のみを用いて要素技術の抽出を行っている．本研究では，要素技術だけでなく，技術の効果も抽出する．また，論文の表題だけでなく，論文と特許の概要を分析の対象にすることで，より網羅性の高い技術動向分析を目指す．

### 3.2 技術動向分析

研究動向の調査に関して，村田ら[村田 2005]の研究がある．村田らは，言語処理学会年次大会および論文誌の論文表題から名詞を抽出し，様々な側面から自然言語処理分野の研究動向の分析を行っている．この分析では，論文表題中の名詞は全て等価に扱われているが，本研究では，論文表題の構造を解析することにより，要素技術と効果を示す用語を識別する．

西山ら[西山 2009]は，技術文書から特定の技術エリアで生み出される新製品・新技術に関する記述をすばやく把握したいというニーズに応える技術文書マイニング手法を提案している．西山らは複数の手掛かり語を用い，ルールベースで効果に関する記述個所を抽出している．しかしながら，抽出に用いる手掛かり語が限定的であるため，例えば，「精度が 0.935」などの数値で表現される効果には対応できない，という問題がある．本研究では，係り受け関係や分布類似度などの統計的手法を用いて半自動的に収集した手掛かり語を用いることにより，多様な効果に関する記述の抽出を目指す．

## 4. 論文および特許概要の構造解析

### 4.1 論文および特許概要の構造解析手法

前述のように，人手で作成したルールに対応付けて表題や概要を解析している研究は少なくない．しかし，様々な表現が存在する概要全てをルールに対応付けて構造解析することは難しい．本研究では，この問題を，「概要の各単語に次節のタグのいずれかを付与」するという系列ラベリング問題として考え，機械学習を用いてタグの自動付与を行う．

### 4.2 タグの定義

以下に，本研究で扱うタグを定義する．

- **TECHNOLOGY**： 要素技術を示す．
- **EFFECT**： 効果（新しい機能の追加，新しく得られた物質，精度などの数値または増加・減少，問題点の抑制や解決したこと，明らかになったこと）を示す．EFFECT タグは，以下に示す ATTRIBUTE タグと VALUE タグを含む．
- **ATTRIBUTE と VALUE**： 例えば，「処理速度(ATTRIBUTE)が向上(VALUE)」や「精度(ATTRIBUTE)が 0.935(VALUE)」のように技術の効果部は「属性(ATTRIBUTE)」と「属性値(VALUE)」の対で表現できる．ATTRIBUTE は，この属性部分を示す．

以下に，「PM 磁束制御用コイルを設けて閉ループフィードバック制御を適用するため，電気損失を最小化できる．」という概要に上記のタグを付与した例を示す．

> PM 磁束制御用コイルを設けて<TECHNOLOGY>閉ループフィードバック制御</TECHNOLOGY>を適用するため，<EFFECT><ATTRIBUTE>電気損失</ATTRIBUTE>を<VALUE>最小化</VALUE></EFFECT>できる．

### 4.3 論文および特許概要の構造解析の基本方針

論文および特許概要中の「を用いた」や「を具備する」といった表現の直前には要素技術を表す用語が出現する．一方で，「が可能になる」や「ができる」の直前には効果を表す用語が出現する可能性が高い．そこで，このような手掛かり語のリストを作成しておき，各々のリスト中の手掛かり語の有無を機械学習の素性として用いる．この他，「精度」や「信頼性」のように属性になりやすい用語や，「向上」や「高速化」のように属性値になりやすい用語が存在する．このような用語を収集してリストを作成しておけば，これらの用語の有無を機械学習の素性として用いることができる．しかし，様々な分野の属性と属性値を人手で網羅的に収集するのは容易ではない．そこで，本研究では，係り受け関係や分布類似度などの統計的な手法を用いて半自動的に手掛かり語リストを作成する．次節では，その収集方法と精度向上のための新たな素性について述べる．

### 4.4 手掛かり語リストの作成

以下に，手掛かり語リストの作成方法について説明する．

・(手順 1) 係り受け関係による収集

特許文書集合から，"向上する"などの属性値になりうる特定の動詞に，"精度(が) "や"効率(を) "などガ格やヲ格で係る名詞/名詞句を属性に関する表現として収集する．

・(手順 2) 上位下位関係による収集

特許文書集合から「A などの効果」という表現を含んだ文を収集し，さらに，A に該当する個所から"改良"や"最適化"などの属性値に関する表現を抽出する．

・(手順 3) 分布類似度による収集

上記の手法で得られた属性および属性値のリストを基に，分布類似度を用いて，新たな属性および属性値に関する表現を収集する．これらの表現を収集する際，あらかじめ，10 年分の特許公開公報約 5 億文を構文解析し，名詞ごとに共起語ベクトル(各名詞と係り受け関係にある動詞を頻度順にまとめた索引語リスト)を作成する．次に汎用連想計算エンジン GETA を用い，リスト中の用語と類似する語を新たな手掛かり語として追加する．なお，この手法で後述の全てのリストを拡張することも可能であるが，予備実験の結果から，属性・属性値の用語リストの拡張のみにおいて，精度が向上することが分かっている．

次に，この他の 2 つの素性について説明する．論文や特許の概要には，主題が記述されている個所がある．このような個所に TECHNOLOGY タグが誤って付与されないように，手掛かり語を用いて判定する．例えば，"提案する"の直前はその論文の主題となりやすく，TECHNOLOGY タグが付与されることはない．そこで，このような手掛かり語の有無を素性のひとつとして用いる．

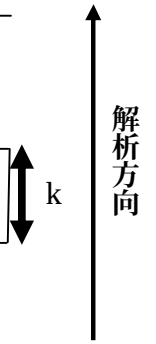
もう 1 つの素性は，特徴的な概要の構造を利用する．論文概要は前半部に研究目的や提案技術が，中間部には要素技術が，後半部にはまとめや効果部が記述されることが多い．また，特許においても【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】という項目で構成されている．そこである文字列が，これらの 3 つの構成部分のどこに属するのかを素性として用いる．

### 4.5 機械学習に用いる素性

概要の構造解析を行う際，機械学習に以下の 9 個の素性を用いる．括弧内の数値は各リストの個数である．

表 1. 機械学習に用いる入力データ(概要例)

| 概要中の各単語 | 品詞 | F1 | F2 | F3 | F4 | F5 | F6 | F7 | タグ |
|---|---|---|---|---|---|---|---|---|---|
| 電気 | 名詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | |
| 損失 | 名詞 | 1 | 0 | 0 | 0 | 0 | 0 | 3 | |
| を | 助詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | |
| 最小 | 名詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | B-VALUE |
| 化 | 名詞 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | I-VALUE |
| でき | 動詞 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | O |
| る | 助動詞 | 0 | 1 | 0 | 0 | 0 | 0 | 3 | O |
| よう | 名詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | O |
| に | 助詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | O |
| なる | 動詞 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | O |




1) 概要中の各単語
2) 品詞情報
3) ATTRIBUTE-internal(1210)
属性の手掛かり語の有無. (例, 処理量, 精度)
4) EFFECT-external(21)
効果部の手掛かり語の有無. (例, できる, 実現する)
5) TECHNOLOGY-external(45)
要素技術の手掛かり語の有無. (例, を用いた, に基づいた)
6) TECHNOLOGY-internal(17)
要素技術専門用語の有無. (例, HMM, SVM)
7) VALUE-internal(408)
属性値の手掛かり語の有無. (例, 増加, 抑止)
8) HEAD-exclusion(12)
主題となる不要語または主題の手掛かり語の有無. (例, を提案, 開発)
9) Location
概要の構造に関する素性. 前半部を"1", 中間部を"2", 後半部を"3"とした.

## 5. 実験

提案手法の有効性を調べるため, 実験を行った. 本節では, まず, 実験で用いたデータやツールを 5.1 節で述べ, 5.2 節でその結果を報告する.

### 5.1 実験データ

・実験データ

NTCIR ワークショップ 8 特許マイニングタスク[Nanba 2010]のデータを用いて実験を行った. このデータは, 1993～2002 年の日本国公開特許公報から任意に選択された 500 件に含まれる 3 つの項目【発明が解決しようとする課題】【課題を解決するための手段】【発明の効果】に TECHNOLOGY, EFFECT, ATTRIBUTE, VALUE タグが人手で付与されている. また, 同一のタグが, 論文概要 500 件に付与されている. このうち, 2010 年 1 月現在, 訓練用データとして公開されている特許 200 件および論文概要 250 件を訓練用データとして, ドライラン用データとして公開されている特許および論文概要, 各 50 件を評価用データとして用いる.

・機械学習に用いるツールと入力データ

概要の構造解析には SVM ベースのチャンキングツールである yamcha を用いる. 機械学習で用いる入力データの例を表 1 に示す. 表において, 1 列目は概要中の単語を, 2 列目は各単語の品詞を示す. 形態素解析には MeCab を用いる. 3 列目以降は, それぞれ, 4.5 節で説明した ATTRIBUTE-internal リスト, EFFECT-external リスト, TECHNOLOGY-external リスト, TECHNOLOGY-internal リスト, VALUE-internal リスト, HEAD-exclusion リストの語の有無を示している. また, 9 列目は Location 素性を示している. 右端の列は教師用データを示す. yamcha は, 表 1 の網掛けで囲まれた個所にタグを付与する場合, 窓幅を k とすると, 前後 k 行の素性と現在の行の素性, 前 k 個のタグを素性として用いる. 予備実験の結果から論文の概要構造解析には窓幅 3 を, 特許の概要解析には窓幅 4 を用いる.

### 5.2 評価実験

論文と特許の概要構造解析それぞれについて評価実験を行った. 評価尺度には精度と再現率を用いた. ベースラインとして, 「概要に含まれる単語」, 「品詞」, 人手で収集した「手掛かり語の有無」(4.4 節における手順 1 と 2 のみ利用)を用いた. 論文の概要構造解析結果を表 2 に示す. また, 特許の概要構造解析結果を表 3 に示す.

表 2. 論文の概要構造解析

| | ベースライン | | 提案手法 | |
|---|---|---|---|---|
| | 再現率 | 精度 | 再現率 | 精度 |
| TECHNOLOGY | 0.171 | 0.481 | 0.171 | 0.500 |
| ATTRIBUTE | 0.090 | 0.545 | 0.119 | 0.533 |
| VALUE | 0.104 | 0.500 | 0.221 | 1.000 |
| Average | 0.178 | 0.539 | 0.217 | 0.595 |

表 3. 特許の概要構造解析

| | ベースライン | | 提案手法 | |
|---|---|---|---|---|
| | 再現率 | 精度 | 再現率 | 精度 |
| TECHNOLOGY | 0.333 | 0.576 | 0.404 | 0.519 |
| ATTRIBUTE | 0.198 | 0.513 | 0.327 | 0.660 |
| VALUE | 0.198 | 0.541 | 0.366 | 0.740 |
| Average | 0.260 | 0.539 | 0.373 | 0.584 |

論文の概要構造解析の結果では, 再現率 0.217, 精度 0.595 が得られた. また, 特許の概要構造解析の結果では, 再現率 0.373, 精度 0.584 が得られた. 論文, 特許共に, VALUE タグの解析精度が最も良かった.

## 6. 考察

まず, 論文の概要に関する構造解析の解析誤りについて 6.1 節で述べ, 次に, 特許の概要構造に関する解析の解析誤りについて 6.2 節で述べる.

### 6.1 論文の概要に関する解析誤り

論文の概要構造解析の解析誤りを回数が多いものから 4 種類を挙げる.

Ⅰ) ATTRIBUTE における, "の", "による"(14%)
Ⅱ) TECHNOLOGY-internal の不足(13%)
Ⅲ) 単位(6%)
Ⅳ) 文脈依存(6%)
以下に, それぞれの解析誤りについて説明する.

**Ⅰ) ATTRIBUTE における"の", "による"**

「指向性の影響を低減」という例では, "指向性の影響"の個所に ATTRIBUTE タグが付与され, "低減"の個所に VALUE タグが付与されるべきであるが, いずれについてもタグが付与されなかった. "影響"には VALUE-internal 素性が, "低減"も同様に VALUE-internal 素性が立っており, 判定できなかったと推測される. 概要の単語が素性に存在するかどうかではなく, VALUE-internal 素性が近接する文においては, VALUE-internal 素性の適用を一考する必要があると考えられる.

**Ⅱ) TECHNOLOGY-internal の不足**

「SAW 素子を用いた」という例では, "SAW 素子"の個所に付与されるべき TECHNOLOGY タグが付与されていない. "を用いた"は TECHNOLOGY-external リストに含まれるが, "SAW 素子"は TECHNOLOGY-internal 素性リストに含まれておらず, 要素技術と判定されなかったと推測される.

**Ⅲ) 単位**

「熱回収量は 77w/m^2」という例では, "熱回収量"の個所に ATTRIBUTE タグが付与され, "77w/m^2"の個所に VALUE タグが付与されるべきであるが, いずれも付与されておらず, 単位付きの数値に VALUE タグが付与されない傾向があると考えられる. 改善策として, 単位リストを VALUE-internal リストに追加する方法がある. 数値直後の名詞や記号の列を収集し, 頻度順に並べることで単位リストの作成を試みたが, 十分な結果が得られなかったため, 別方法での収集が必要となる.

**Ⅳ) 文脈依存**

「N-gram モデルを用いて」という例では, "を用いて"が TECHNOLOGY-external リストに存在するため "N-gram モデル"が要素技術として判定される. しかし, 同じ概要中で「N-gram モデルだけでは」という文が記載されている場合, 1 度要素技術と判定された語が出現しても, "だけでは"が TECHNOLOGY-external リストに存在しないため, 要素技術として判定されない問題がある. この問題は, 機械学習の後処理として, 要素技術を一旦格納して, タグを振り直すという 2 段階の手法により解析精度が改善される可能性がある.

### 6.2 特許の概要に関する解析誤り

特許の概要構造解析の解析誤りを回数が多いものから 4 種類を挙げる.

ⅰ) 論文独特な表現方法(33%)
ⅱ) 文脈依存(16%)
ⅲ) ATTRIBUTE の"の", "による"(7%)
ⅳ) ATTRIBUTE, VALUE の出現順(7%)

特許の概要構造解析の誤りも, 論文と同じ誤りが確認された. ⅱとⅲは 6.1 節で述べたので, 本節では, ⅰとⅳの解析誤りについて述べる.

**ⅰ) 論文独特な表現方法**

特許の要素技術部は, 論文に記述されているような端的なものではなく, 技術を一般的な表現で長く記述するという特徴がある. 例えば, 「位相シフトを行う前後のレベルを比較し, 増加すれば加える位相シフトの方向を維持し, 減少すれば位相シフトの方向を逆転する手段」などが要素技術の例として挙げられる. また, 概要の構造は「(要素技術 A)と, (要素技術 B)と, (要素技術 C)とを設け」という定型であること多い. そのため, 「該遮断手段を作用状態から非作用状態に設定し, その後操作されても非作用状態を保持するという操作手段とを設け」という例の場合, "該遮断手段を作用状態から非作用状態に設定し, その後操作されても非作用状態を保持するという操作手段"の個所が 1 つの要素技術としてタグ付与されるべきであるが, "設定し, "の読点直後の語である, "その後操作されても非作用状態を保持するという操作手段"が誤って要素技術として判定されている.

**ⅳ) ATTRIBUTE, VALUE の出現順**

「高い認識率」という例では, "高い"の個所に VALUE タグが付与され, "認識率"の個所に ATTRIBUTE タグが付与されるべきであるが, いずれのタグも付与されなかった. 機械学習をする際のトレーニングデータには「精度が高い」のように ATTRIBUTE, VALUE となる単語の並びが多い. そのため, VALUE, ATTRIBUTE の順番で単語が出現した場合, "高い"より前の語に ATTRIBUTE が存在しないか, もしくは"認識率"の後ろの語に VALUE がないかと判断しているためタグ付与ができなかったということが考えられる.

## 7. おわりに

本研究では, 特定分野の特許概要と論文概要から, 要素技術とその効果を示す表現を自動的に抽出し, 論文と特許を, 「要素技術」と「効果」という 2 つの観点で分類した. その結果, 論文の概要構造解析では, 0.217 の再現率, 0.595 の精度が得られ, 特許の概要構造解析では, 0.373 の再現率, 0.584 の精度が得られた.


## 謝辞

本研究で用いた論文と特許のデータは, 国立情報科学研究所の許可を得て, NTCIR テストコレクションを利用させていただいた.


## 参考文献


[近藤 2007] 近藤, 難波, 奥村, 新森, 谷川, 鈴木 : "論文データベースからの研究動向情報の抽出". 言語処理学会第 13 回年次大会, pp.470-473, 2007.

[近藤 2008] 近藤, 難波, 竹澤 : "翻訳知識を用いた英語論文表題の構造解析". 自然言語処理研究会, NL-187, pp37-43, 2008.

[村田 2005] 村田, 一井, 馬, 白土, 井佐原 : "過去 10 年間の言語処理学会論文誌･年次大会における研究動向調査". 言語処理学会 11 回年次大会, 2005.

[西山 2009] 西山, 竹内, 渡辺, 那須川, 武田 : "新技術が持つ特長に注目した技術調査支援ツール". 人工知能学会論文誌, Vol24 No.6 pp.541-548. 2009.

[Nanba 2010] Nanba,H., Fujii,A., Iwayama,M., and Hashimoto,T. "Overview of the Patent Mining Task at the NTCIR-8 Workshop". Proceedings of the 8th NTCIR Workshop Meeting, 2010.