技術文書からの動向情報の抽出と可視化

福田 悟志　難波 英嗣　竹澤 寿幸

広島市立大学 情報科学部

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，2010 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 2010」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，本研究では，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を目指す．

特定分野の技術動向を効率的に把握するためのこれまでの研究として近藤ら[近藤 2010]の研究がある．近藤らは，要素技術とその効果を示す表現を自動的に抽出し，「要素技術」と「効果」という 2 つの観点から論文と特許を分類している．しかし，効果に関する表現において，近藤らの手法では考えられるすべての素性を考慮していないため，十分な抽出結果が得られていない．そこで本研究では，近藤らの手法に新たな素性を追加することにより，精度改善を目指す．また，近藤らは，論文用および特許用の抽出器を機械学習により獲得するため，論文用および特許用タグ付きコーパスを単独で用いていたが，本研究では，ドメイン適応技術のひとつである FEDA を用い，ドメインの異なる 2 種類のコーパスを用いた機械学習を試みる．

本論文の構成は以下の通りである．次節では，関連研究について述べ，3 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．4 節では，有効性を調べるために行った実験について述べ，結果について考察する．最後に 5 節で本稿をまとめる．

2.関連研究

●単位素性を用いた情報の抽出

Nishiyama ら [Nishiyama 2010]は，日本語論文，日本語特許内で記述されている各単語を様々な字種として区別し，要素技術と効果部の抽出を効率的に行なっている．平仮名，片仮名，英字，数字，単位，および記号をクラスとして定義し，各単語を 1 つ以上のクラスに分類することで，論文と特許の構造解析を効果的に行っている．本研究では，クラスとして定義された単位に着目し，半自動的に収集した数値付き単位を手がかり語の素性として機械学習に用いる．

●ドメイン適応を用いた情報抽出

FEDA とは，元ドメインのデータを併用して，目標ドメインの性能を改善するドメイン適応技術である[Daumé III 2007]．FEDA は，元ドメインの特徴ベクトルを$x^{(S)}$，目標ドメインの，特徴ベクトルを$x^{(T)}$とした時，元ドメインのデータでは$\langle x^{(S)}, x^{(S)}, 0\rangle$のようなベクトル，目標ドメインでは$\langle x^{(T)}, 0, x^{(T)}\rangle$のように，長さが 3 倍の高次元の特徴ベクトルに変換を行う．そして，変換後の特徴ベクトルを用いて通常の方法で学習を行う．この手法により，従来のドメイン適応手法とほぼ同程度の精度結果が得られることが明らかにされている．また，複数のドメインを用いる場合，ベクトルをドメイン数+1 の次元に拡張すれば対応出来ることも特徴である．

Nishiyama らは，日本語論文および日本語特許を対象として FEDA を用いることにより，解析精度が向上することを示している．本研究でも同様に，FEDA を用いてその有効性を確認する．

3. 論文および特許概要の構造解析

3.1 論文および特許概要の構造解析手法

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，概要には様々な表現が存在し，その全てをルールに対応付け，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目指す．

3.2 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．

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．近藤ら[近藤

2010]は，このような手掛かり語リストを作成し，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た．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の収集し，リストの作成も行った．しかし，1節でも述べたように，数値や数値付き単位となるものに対して十分な抽出結果を得られていない．そこで本研究では，これらの手掛かり語リストに加え，単位リストの作成を行った．次節では，その収集方法について述べる．

3.3 単位リストの作成

以下に，単位リストの作成方法について説明する．

・(手順 1) 数値付き単語の収集

直前に数値が記述されている単語は単位になりうる場合が多い．そこで，論文文書集合から，"100cm"や"10kg"などのように直前に数値が記述されている単語の収集を行う．

・(手順 2) 単語の選別

上記の手法で収集した単語には，" . " や ") "など明らかに単位ではない単語も収集されている．また，収集した単語の中において，頻度の少ない単語は単位に関係ないものが多く含まれていると考えられる．そこで，収集した単語 4927 語において，単位になりうる単語を人手で選定し，単位リストを作成する．

また，本研究では日本語論文だけでなく，日本語特許の解析も行う．日本語論文では英字や記号を半角で記述するが，日本語特許では英字や記号を全角で記述する傾向がある．日本語特許の解析を行うために，作成した単位リストから半角英字や半角記号を対象に，それぞれの全角英字，全角記号を単位リストに加える．

3.4 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 10 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．

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．
10)UNITS-internal (274)
数値付き単位の有無．(例，88%，30 種)

3.5 FEDA の応用

Nishiyama ら[Nishiyama 2010]の手法では，論文ドメインと特許ドメインの素性を単純に混ぜあわせた素性を学習に用いる事によって，解析精度の向上を図った．本研究では，単純に論文ドメインと特許ドメインの素性を混ぜあわせて学習を行っていくだけでなく，FEDA を応用した形で，様々な方法を提案し学習に用いる．

一般に，日本語特許は日本語論文に比べ，一文が長く記述されている．ゆえに，日本語特許に付与される構造タグも日本語論文に比べ，長く付与される傾向がある．特に，要素技術に対する構造タグの付与について，例えば，日本語論文における構造タグの付与は，「<TECHNOLOGY>X 線回折パターン</TECHNOLOGY>」と付与されるが，日本語特許では「<TECHNOLOGY>前記第 1 電極パターンを被覆するように設けられた高抵抗値の第 2 電極パターン</TECHNOLOGY>」と長く付与される傾向がある．そこで本研究では，「技術文書 A と技術文書 B のドメインの素性を混ぜあわせて機械学習を行う」，「技術文書 A と要素技術関連の素性を除いた技術文書 B のドメインの素性を混ぜあわせて機械学習を行う」，という 2 つの手法を論文および特許の概要構造解析に用いる．

また，論文ドメインと特許ドメインの素性を用いて構造解析を行う場合，それぞれを混ぜあわせて学習に用いるだけでなく，あるドメインの素性で一旦学習させタグの付与を行った後，もうひとつのドメインの素性を用いて学習を行い，タグの付与を行っていくという方法が考えられる．そこで，上記の手法に加え，「技術文書 A のドメインの素性を用いて一旦機械学習を行いタグの付与を行った後，技術文書 B のドメインの素性を用いて機械学習を行う」，「技術文書 A のドメインの素性を用いて一旦機械学習を行いタグの付与を行った後，要素技術関連の素性を除いた技術文書 B のドメインの素性を用いて機械学習を行う」，という 2 つの手法も用いて論文および特許の概要構造解析を行う．

4. 実験

3 節で述べた提案手法の有効性を調べるため，評価実験を行った．実験で用いたデータやツールを 4.1 節で述べ，4.2 節で比較手法について述べる．4.3 節で解析結果について述べ，4.4 節で考察を述べる．

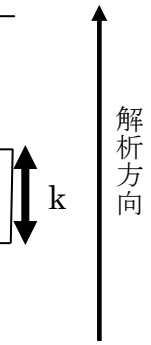
4.1 実験データ

・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，300 件を訓練用データ，200 件を評価用データとして用いる．

表 1. 機械学習に用いる入力データ(概要解析の例)

概要中の各単語	品詞	F1	F2	F3	F4	F5	F6	F7	F8	タグ
電気	名詞	0	0	0	0	0	0	3	0	
損失	名詞	1	0	0	0	0	0	3	0	
を	助詞	0	0	0	0	0	0	3	0	
最小	名詞	0	0	0	0	0	0	3	0	B-VALUE
化	名詞	0	0	0	0	1	0	3	0	I-VALUE
でき	動詞	0	1	0	0	0	0	3	0	O
る	助動詞	0	1	0	0	0	0	3	0	O
よう	名詞	0	0	0	0	0	0	3	0	O
に	助詞	0	0	0	0	0	0	3	0	O
なる	動詞	0	0	0	0	0	0	3	0	O



・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. 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 素性を示しており, 10 列目は UNITS-internal リストの語の有無を示している. 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.

4.2 比較手法

以下に, 論文, 特許の構造解析の際に用いる比較手法を述べる.

・BASELINE : 3.4 節で述べた素性において,「数値付き単位の有無」以外の素性を用いて機械学習を行う.
・METHOD1 : 3.4 節で述べた素性を全て用いて機械学習を行う.
・METHOD2 : METHOD1 に加え,「技術文書 A と技術文書 B のドメインの素性を混ぜあわせた素性」を用いて機械学習を行う.
・METHOD3 : METHOD1 に加え,「技術文書 A と要素技術関連の素性を除いた技術文書 B のドメインの素性を混ぜあわせた素性」を用いて機械学習を行う.
・METHOD4 : METHOD1 に加え,「技術文書 A のドメインの素性を用いて一旦機械学習を行いタグの付与を行った後, 技術文書 B のドメインの素性を用いて機械学習を行う」.
・METHOD5 : METHOD1 に加え,「技術文書 A のドメインの素性を用いて一旦機械学習を行いタグの付与を行った後, 要素技術関連の素性を除いた技術文書 B のドメインの素性を用いて機械学習を行う」.

4.3 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, および F 値を用い, AVERAGE の算出には, NTCIR-8 に合わせるために, 表題および概要の TECHNOLOGY タグに加え, 概要の ATTRIBUTE タグと VALUE タグを用いた. また, 表題の構造解析結果は, 近藤ら[近藤 2010] の表題構造解析結果を用いた.

比較手法の評価結果の AVERAGE を表 2 に示す.

表 2. 論文の表題および概要の構造解析結果

	論文		
	再現率	精度	F 値
BASELINE	0.184	0.686	0.290
METHOD1	0.191	0.669	0.298
METHOD2	0.211	0.547	0.305
METHOD3	0.211	0.596	0.311
METHOD4	0.246	0.411	0.308
METHOD5	0.254	0.496	0.336

表 3. 特許の表題および概要の構造解析結果

	特許		
	再現率	精度	F 値
BASELINE	0.441	0.537	0.485
METHOD1	0.441	0.537	0.484
METHOD2	0.429	0.540	0.478
METHOD3	0.426	0.552	0.481
METHOD4	0.454	0.493	0.473
METHOD5	0.455	0.507	0.480

4.4 考察

以下では, 論文, 特許の概要構造解析において, 提案手法を用いた際に起こった解析誤りについて述べる.

●単位リストを用いた手法に関する考察

論文の概要構造解析において,「従来受理出来なかった入力の 40%以上を正しく認識することが可能になること」という例では, " 従来受理出来なかった入力の 40%以上"の個所に ATTRIBUTE タグが付与され, " 認識"の個所に VALUE タグが付与されるべきであるが"探索空間"と"増加"に ATTRIBUTE タグが付与され,"40%以上"に VALUE タグが付与された. BASELINE では, 同様の例に対していずれについてもタグが付与されていなかったことから, 単位リストに含まれている"%"によって"40%以上"の個所に VALUE タグが付与されたと推測される.

特許の概要構造解析において, AVERGE の F 値が低下した事から, 提案手法の有効性が得られなかった. これは, 論文では学術的な研究成果を理論的に述べるために, データの詳細な結果や解釈の記述を行う

のに対し，特許では新規性があり且つ有用な発明について詳細に記述を行うので，数値付き単位が特許文中に出現する頻度は少ないからと考えられる．

●FEDA に関する考察

論文，特許の概要構造解析において，それぞれの手法に共通する解析誤りの原因を考察する．

1 つ目の原因として，論文と特許のドメインの素性を用いたことで，学習を行うための情報量が増えたため，付与されたタグの数が異なったことが考えられる．特許における METHOD2,3 の手法では，付与されたタグの数はおよそ 100 個減少した．これは，論文と特許のドメインの素性を混ぜ合わせているため，論文ドメインの素性が特許ドメインの素性による情報をある程度抑えているからと考えられる．これにより，AVERAGE における精度は向上したが，正しいタグの付与の情報も抑えてしまっているため，再現率は低下している．一方で，論文における METHOD2,3,4,5 の手法および特許における METHOD4,5 の手法では，100 以上の数のタグが付与された．これにより，正しい部分にタグの付与も行われたが，間違った部分にタグの付与も多く行われたため，再現率は向上しているが，精度は低下している．

2 つ目の原因として，構造タグが不当に付与されていることが挙げられる．これは主に，一旦構造タグを付与した技術文書にさらに構造タグを付与する手法を用いている MERHOD4,5 に現れた．論文において，「87.4%(特定話者:92.1%)の音素認識率を得た」という例では，論文ドメインの素性を用いて"音素認識率"に ATTRIBUTE タグが付与されたが，特許ドメインの素性を用いてタグの付与を行った時，先程付与した構造タグを含め，" 87.4%(特定話者 : 92.1%)の<ATTRIBUTE>音素認識率</ATTRIBUTE>"に付与を行っている．また，特許において，「磁気抵抗効果を有しかつバイアス磁界が付与されると共に該磁気抵抗効果を用いて情報を再生するための再生素子」という例では，特許ドメインの素性を用いて" 磁気抵抗効果を有しかつバイアス磁界が付与されると共に該磁気抵抗効果を用いて情報を再生するための再生素子"に TECHNOLOGY タグの付与を行ったが，論文ドメインの素性を用いてタグの付与を行った時に，先程付与した構造タグにおいて ATTRIBUTE タグは" <ATTRIBUTE><TECHNOLOGY>磁気</ATTRIBUTE>"と付与され，VALUE タグは" <VALUE>効果</VALUE>"と付与されている．

また，論文の解析誤りに特に現れた 3 つ目の原因として，日本語特許は日本語論文に比べ，一文が長く記述されているために，必要以上に長く構造タグを付与されてしまう事が挙げられる．「完全なブロック構造を持たない帯域分割手段との整合性に優れ」という例では" 整合性"に ATTRIBUTE タグが付与され，"優れ"に VALUE タグが付与されるべきであるが，VALUE タグは正しく付与されていたが，ATTRIBUTE タグは"帯域分割手段との整合性"に付与された．

5.おわりに

特定分野の特許および論文概要から，要素技術とその効果を示す表現を効果的に抽出するために，単位リストの作成をした．その結果，論文の構造解析では，有効性が確認されたが，特許の構造解析では，従来手法とほぼ値に変化が無かったため，あまり効果がないことが確認された．また，論文または特許概要の構造解析を行う際に，論文と特許それぞれの素性を様々な方法で組み合わせることによって解析精度の向上を試みた．その結果，論文の構造解析では，「論文ドメインの素性を用いて一旦機械学習を行いタグの付与を行った後，要素技術関連の素性を除いた特許ドメインの素性を用いて機械学習を行う」という手法が最も効果的であり，提案手法の有効性が確認されたが，特許の構造解析では，様々な手法を用いたが，十分な精度結果は得られなかった．

今後の課題として，単位リストに含まれている単語の選定と，論文，特許ドメインの素性を用いる際に，不当なタグの付与を行わないための処理方法の提案および必要とされる素性情報の選定が挙げられる．また，係り受け関係による構文解析を行うことで，より網羅的に文の構造を把握させれば，不必要な部分への構造タグの付与や，構造タグの長さが必要より長くまたは短く付与されたという解析誤りを改善できる可能性がある．

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 2010] 近藤，難波，竹澤 ："特許と論文からの技術動向マップの自動構築". 言語処理学会第 16 回年次大会, 2010.

[Nishiyama 2010] Risa N, Yuta T, Yuya U, and Hironori T. "Feature-Rich Information Extraction for the Technical Trend-Map Creation". Proceedings of the 8th NTCIR Workshop Meeting, 2010.

[Daumé III 2007] Daumé III. "Frustratingly Easy Domain Adaptation". Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 256-263, 2007.

[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.