3B3-NFC-4-4

技術文書の情報編纂: 課題・特長・手段を表す表現の抽出と利用

Information Compilation for Technical Documents: Extraction and Utilization of Phrases Mentioning Issues, Advantages and Methods

西山 莉紗
Nishiyama, Risa

日本アイ・ビー・エム株式会社　東京基礎研究所
IBM Research - Tokyo

The author and her colleagues have presented their practices of extracting three types of phrases from technical documents, particularly from patent documents: issues to be solved by the new technology, advantages of the technology, and methods used to solve the issue. These phrases are important to understand the patent documents and are expected to be utilized for mining these documents. This work is one example of the Information Compilation Challenge, which is aiming at improving the quality of information access systems by utilizing natural language processing methods and visualization methods. This paper introduces and summarizes a series of work by authors and discusses future work.

1. はじめに

本稿では情報編纂の基盤技術チャレンジの一環として著者らがこれまでに取り組んできた，技術文書からの三種類の表現の抽出と，その技術文書マイニングへの利用について概要と成果を紹介する．ここで言う技術文書とは，科学技術論文や各社が出しているホワイトペーパー，プレスリリース，そして公開特許公報などの，新技術の効果や特徴について述べた文書を指す．

情報編纂の基盤技術では様々な構造・非構造情報を集約し，利用者の多様な興味に合わせて提示する技術の発展と整理を目指している [加藤 06]．情報編纂を実現する上で，自然言語処理技術を用いて文書からある特定の意味を持った表現を自動的に抽出することや，文書検索結果の順序付けにその抽出結果を利用することは重要な基盤技術の一つになると考えられる．本研究では特に，企業の開発製造部門担当やコンサルタントまたは技術者などの技術情報アクセスを支援することを目的とした技術文書マイニングをとりあげ，そこでの自然言語処理技術の利用可能性を示すことを目標としてきた．

技術文書マイニングに役立つことが期待される表現として，著者らはこれまでに以下の三種類の表現の抽出に取り組んできた．

- 課題表現: 「コストが高い」,「装置が大きくなる」のような当該領域において解決されることが望まれる不具合や障壁などを示す表現
- 特長表現: 「コストを低減する」,「装置が小さくなる」のような，当該技術の長所を示した表現．
- 手段表現: 「隠れマルコフモデル」,「フィードバック制御」などの，課題の解決に用いられた手法や物質の名前を表す表現

課題表現と特長表現の関係は，評価表現抽出として主にレビューテキストを対象に研究が進められている，好評（ポジティブ）・不評（ネガティブ）表現とも関連が深い [Pang 08, 乾 06]．レビューテキストの分析では，ある表現がポジティブとネガティブのどちらの極性に属するかということは，レビューの対象となっている商品の分野に大きく依存することが指摘されている（この問題は一般に分野依存性の問題と呼ばれる）．同様に，技術文書からの課題・特長表現抽出も分野依存性の問題を持っている．評価表現抽出では例えば小林ら [小林 05] によって整理された評価表現辞書など，既知のポジティブ・ネガティブ表現を集めた言語資源を利用してそのような分野依存性の問題を解決することが多い．しかし，レビューテキストなどの書き手が自身の意見を中心に書いた文書に対し，本研究が扱っている技術文書では客観的に技術の課題と特長が述べられるため，既存研究で用いられてきた言語資源をそのまま利用することが難しく，別の方法を利用して抽出する必要がある．

本稿ではまず課題・特長・手段表現の抽出方法を説明し，次に，抽出した表現の技術文書マイニングへの利用方法を紹介する．最後に，技術文書マイニングとそこでの自然言語処理技術の活用に関する今後の展望をまとめる．

連絡先: 西山 莉紗，日本アイ・ビー・エム株式会社 東京基礎研究所，lisa@jp.ibm.com，http://www.research.ibm.com/trl/people/lisa/

2. 課題・特長・手段表現の抽出

課題・特長・手段表現の抽出には，これまでに (1) 抽出する表現に現れやすい動詞や形容詞（用言）を中心に利用したルールベースの抽出手法 [西山 09]，(2) 抽出する表現の後に現れやすい文末表現を利用した抽出方法 [西山 10a]，そして (3) 系列ラベリング手法を用いて，教師データから抽出する表現そのものやその周辺の単語の特徴を学習して抽出する方法 [Nishiyama 10] の三種類の方法を試みてきた．本節ではこれらの抽出方法について簡単に説明する．なお，詳細については個々の参考文献を参照していただきたい．

2.1 抽出の対象とした技術文書

抽出の対象となる技術文書として，著者らは主に公開特許公報を扱ってきた．その理由の一つは，特許が持つ重要性と速報性である．特に製造業にとって，新技術の特許出願を行うことは自社の技術を保護する上で必要不可欠である．そのため，公開特許公報には新技術の情報が網羅的に掲載されていることが期待される．また，論文発表やプレスリリースなどで一度公のものとなった技術については特許を取得することができなくなることから，他の技術文書と比較して，公開特許公報に

は最も早く技術情報が記載されることが期待できる．ただし，公開特許公報の短所として，出願から公開されるまでに 2 年かかることが挙げられるが，技術の製品化に先んじて特許が出願される事情を鑑みると，公開までの期間を差し引いてもなお新規な技術が書かれていることを期待してよいだろう．

公開特許公報を用いる第二の理由として，文書の様式が特許法で定められているため，セクションの見出しと内容が全ての文書間で共通しており，高い精度で情報抽出を行えることが期待できるためである．このことを利用し，本研究では「発明が解決しようとする課題」という，従来技術が抱えていた課題を中心に記述するセクションから課題表現を,「発明の効果」という，発明の長所について記述するセクションから特長表現を抽出する．なお，手段表現については，上記二セクションと,「問題点を解決するための手段」という発明の構成について書いたセクションの，合わせて三セクションから抽出する．このようなセクションの情報を抽出に用いることができるのは公開特許公報独自の特徴であるが，以下に示す表現の特性を用いた抽出方法は他の技術文書でも有効に働くことが期待される．実際に，2.2 小節で述べる構文パターンを利用した抽出方法については，新製品発表のテキストに対しても人間と同じくらいの抽出精度を達成できていることが示されている [西山 09]．また，2.3 小節に述べる手法で公開特許公報から獲得される，各技術分野における課題表現の知識は，他の技術文書に対しても適用可能であると考えられる．

2.2 用言を中心とした構文パターンの利用

特長表現には「向上する」「高める」などの物事の望ましい側面をより伸ばす意味の用言や,「防止する」「抑制する」などの望ましくない側面を押さえこむ意味の用言が多く用いられる．このような特徴を利用し，表 1 にある，用言を中心とした構文パターンを利用して文書中の特長表現を同定し，そしてパターンに合致した箇所から，あらかじめ定めた数単語分係り受け構造を遡った部分までを特長表現として抽出した [西山 09].

ここに挙げた構文パターンは必ずしも網羅的なものではないが，公開特許公報で分野に関わらず広く用いられる表現を包含している．実際に，この構文パターンを利用することで公開特許公報から F 値 7 割程度で特長表現を抽出することができた．しかし，網羅性を上げるためには，より多くのパターンを収集する必要がある．

このとき，例えば表 1 中の「〜できる」という表現は,「〜を向上できる」,「〜を防止できる」のように，他のパターンを含む文の文末表現として現れやすいという特徴がある．言い換えると,「〜できる」は特長を述べる文脈を形成していると言える．次小節ではこのような文脈を形成する表現を利用して，より多くの表現を収集する方法を説明する．

2.3 文末表現の利用

前小節では特長表現と「〜できる」という文末表現の関係について説明したが，課題表現にも同様の文末表現が存在する．例えば「アーチファクトが発生する」という表現は X 線検査装置にとって好ましくない，解決されるべき技術課題を示している．文書中からこのような表現を課題表現として抽出するためには，前小節で示した構文パターンのような，何らかの言語知識が必要となるが，このような表現は「半導体装置の製造コストが上がるという深刻な問題があった」「アーチファクトが発生してしまう」というように,「〜という問題があった」や「〜してしまう」のような文末表現とともに現れることが多い．このような，記述内容が当該技術領域で望ましくない，解決されるべき事柄であることを示唆する文末表現を課題文脈パターンと呼び，課題表現の抽出に利用した [西山 10a].

表 1: 特長表現同定に利用した構文パターンと抽出される表現の例 [西山 09]

構文パターン	抽出される表現例
〜[助詞]+向上する	ユーザの使い勝手を向上する
〜[助詞]+高める	光の利用効率を高める
〜[助詞]+優れる	冷熱サイクル性に優れる
〜可能+[助詞]+なる	強度を確保することが可能となる
〜[動詞]+できる	円滑な空気の流れを確保できる
〜[*]+実現する	回路の安定動作を実現する
〜[*]+できる	正確なキャリブレーションを行うことができる
〜[助詞]+防止する	画像の劣化を防止する
〜[助詞]+抑制する	変動による影響を抑制する
〜[助詞]+低減する	消費電力を低減する
〜不要+[助詞]+なる	再教育が不要となる
〜必要+[助詞]+ない	手作業で試行錯誤的に作成する必要がなくなる
〜こと+[助詞]+ない	転倒するようなことがない

表 2: 課題表現の獲得に用いた課題文脈パターン（カッコ内はふりがなを表す）[西山 10a]

て+しまう，という+問題+が+ある，恐れ+が+ある， 問題点+が+ある，といった+問題+が+ある， 欠点+が+ある，虞れ（おそれ）+が+ある，

課題表現の抽出に当たっては，まず，ほぼ確実に課題表現を伴う課題文脈パターンを数種類用意した．実験では表 2 にある 7 種類を用意した．そして，課題文脈パターンに係る「名詞句+助詞+動詞」（「コストが上がる」など）,「名詞句+助詞+形容詞」（「検出精度が低い」など）,「名詞句+助詞+動詞+助動詞」（「生産性が上がらない」など）を課題表現として抽出した．このとき，課題文脈パターンを一度伴って現れた表現のみを抽出するだけでなく，文書集合中で課題文脈パターンに係りやすい表現については,「コストが上がる。」のように課題文脈パターンを伴わずに出現した場合についても，課題表現として抽出することとした．その結果，課題文脈パターンを伴って現れた表現のみを抽出した場合と比較して，適合率を 7 割程度に保ったまま 2 倍の再現率の課題表現を抽出することができた．

2.4 系列ラベリング手法の利用

著者らはそれまでの課題・特長表現抽出の経験を基に，NTCIR-8 の特許マイニングタスクにおける技術動向マップ作成サブタスクに参加した [Nanba 10]．このタスクでは，将来的に論文や特許を解決手段と効果を軸にしてまとめ上げることを目的として，科学技術論文と公開特許公報を対象とした特長表現と手段表現の抽出に取り組んだ．

このとき扱った特長表現抽出タスクは，例えば「コストを削減する」という表現を抽出するだけでなく,「コスト」が属性（Attribute）,「削減する」が値（Value）に関する表現であることを特定する必要があるという点で，これまでに述べてきた抽出タスクと異なる．また,「コストを削減する」のような

表 3: 系列ラベリング手法を用いた手段表現と特長表現の抽出に用いられた特徴量 [Nishiyama 10]

種別	概要
単語情報	語幹，品詞タグ，文字種,「高」「活」などの接頭辞の有無,「化」「倍」などの接尾辞の有無
文書構造	出現したセクションの種類，セクション中での相対位置
技術分野	文書に割り当てられている IPC コード[*1]
特長表現としての出現しやすさ	特許文書集合において「ことができる」という表現を伴いやすい表現か否か
係り受け情報	同一文中の他の文節への統計的な係りやすさ

自然言語表現だけでなく,「9 割の精度を実現する」(「9 割」が Value,「精度」が Attribute）のような数値を含む表現も特長表現として抽出する必要がある点も異なった．

このタスクにおいては，人名や地名などの固有名詞を抽出する際に利用されている手法を応用して手段・特長表現を抽出することを試みた．2.2 小節と 2.3 小節でこれまでに見てきたように，課題表現と特長表現の抽出にあたっては，抽出すべき表現の周辺にある記述が参考になる場合が多い．手段表現も同様の性質を持つ．例えば，手段表現は「〜を用いた」「〜による」などの表現を伴って書かれることが多い．また，今回のタスクにおいては，例えば Attribute の後に Value が書かれやすいなど，表現同士の関係も参考にできると考えられる．以上で述べたような特徴は，固有名詞抽出手法で既に活用されているため，固有名詞抽出手法で行われるように，各単語を特徴ベクトルで現して，それらの単語が抽出対象の表現の開始位置にあるか（B），中にあるか（I），外にあるか（O）ということを示す BIO タグ [Tjong Kim Sang 99] を，系列ラベリング手法の一つである条件付き確率場（CRF; Conditional Random Fields）を利用して推定するという方法を取った [Nishiyama 10]．

利用した特徴量は表 3 の通りである．これらの特徴量のうち，文書構造を利用した特徴量と，係り受け情報を利用した特徴量が精度の向上に特に寄与した．

次節では，本節で説明した一連の手法によって抽出された表現が技術情報へのアクセスにどのように用いられるかについて説明する．

3. 技術文書マイニングへの利用

技術文書から抽出した課題・特長・手段を表す表現を利用して，著者らはこれまで以下のような技術文書への情報アクセス手法を提案してきた．

1. 新規な技術応用を発見するための技術調査支援ツール
2. 特定の課題を解決可能な技術の検索ツール

以下にそれぞれ説明する．

3.1 新規な技術応用を発見するための調査支援ツール

新技術を利用した価値の高い新ビジネスを検討する際に有力な方策の一つとなるのは，ある技術の新しい応用先を考えることである．実際に，Strategic Capability Network [Bagchi 00] というビジネス戦略分析手法では，技術の応用可能性を可能な限り列挙して，技術とビジネス応用の関係を整理する．

このような背景に基づき，著者らは技術文書から特にある技術分野における新規な応用について言及している可能性が高い特長表現を取り出し，リストの形式で一覧可能な技術文書マイニングツール，CAPHMIT（CApability PHrase MIning Tool）を提案した [西山 09]．

このツールでは，前節で説明した手法によって抽出された特長表現のうち，検索クエリーとして指定された技術分野の文書に出現しにくく，かつ新聞などの技術文書でない一般的な文書に出現しやすい名詞を多く含むものは，新規な技術応用を示す可能性が高いとして，リストの上位に配する．分析者はリストの一覧を見ることで，データマイニング関連技術の応用を概観するだけでなく，各特長表現は元の技術文書の検索スニペットとなっているため，クリックすることで詳細を確認することも可能である．例として，データマイニング分野で新規な応用を示しているとツールによって判断された特長表現の上位 15 位を表 4 に示す．なお，表中で右側に「*」マークが付いている表現は，実際に実験の評価者によって新規な応用を示唆している可能性が高いと判断されたものである．

3.2 特定の技術課題を解決可能な技術の検索ツール

ビジネス上重要な技術課題の解決につながる新技術を把握することは，企業の技術戦略を立案する上で非常に重要である．技術文書集合から特定の課題の解決につながる技術を検索可能にしたり，または同じ効果を持つ技術を集約してユーザーに提示することは，技術動向の把握に大変役立つことが期待される．

このような情報アクセスを可能にするためには，様々な表現で示される特許の効果を認識する必要がある．例えばある情報処理システムにおいて「操作性が悪い」という技術課題を考えたとき,「操作性を向上することができる」を特長とした技術は直接的にこの課題を解決または和らげていると言える．

著者らはある課題表現と解決関係にある複数の特長表現を，特許明細書中の課題表現と特長表現の共起関係を用いて検出することを試みた [西山 10b]．その結果，表 5 にあるように,「信頼性が低下する」と「エロージョンを抑制する[*2]」のように，文字列上は自明ではない解決関係を発見することができた．検索ツール自体はまだ実装されていないが，このような手法で獲得された課題表現と特長表現の対は，ツールを有効に働かせる言語知識として役立つことが期待される．

4. 今後の展望

今後の研究方針として，以下の二点が挙げられる．

1. 技術文書マイニングシステムへの課題・特長・手段の三種類の表現抽出結果のさらなる利用
2. 上記システムを利用することによる情報アクセス改善の評価ならびに評価結果に基づく改良

著者らはこれまで技術情報へのアクセスに利用することを目的として，技術文書から手段・課題・特長という三種類の重要表現を自動的に抽出する方法を中心に検討してきた．一方で，実際にこれらの表現の抽出結果を利用して実装したシステムは 3.1 小節で示した新規な技術応用を発見するための技術調査支

*2 エロージョンとは材料の表面に生じる侵食のことであり，半導体に生じる不良の一つである．

表 4: データマイニング分野において新規な技術応用と判断された特長表現（上位 15 件）[西山 09]

特長表現	スコア	表現中の名詞	
文字入力能力の低い携帯端末などでの利便性が向上する	2.81	能力, 文字入力, 利便性, 携帯端末	*
情報検索精度の低下を防止できる	2.01	低下, 情報検索, 精度	*
種々の情報提供サービスを受ける際の利便性を向上する	1.81	種々, 利便性, 情報提供サービス	*
計算の途中で動的に変更できる	1.80	途中, 計算	
潜在ターゲットを導出できる	1.80	ターゲット, 導出, 潜在	
配達の手配の要求をすることができる	1.73	要求, 配達, 手配	
看護師や環境との関係を的確に把握できる	1.63	環境, 関係, 看護	
処理量削減を実現できる	1.58	削減, 処理量	
プロセス－品質モデルを作成することができる	1.53	品質, プロセス, モデル	*
グラフィックス・イメージを生成することができる	1.41	イメージ, グラフィックス	*
対象品の品質の推測に用いることのできる	1.30	品質, 推測	
送付忘れを防止することができる	1.29	忘れ, 送付	*
リアルタイムに対応するビジネスのスキームを構築することができる	1.17	ビジネス, リアルタイム, スキーム	*
情報処理量の増大を有益に提供することができる	1.16	増大, 情報処理	
現象の変化の詳細な様子を観察することもできる	1.16	変化, 現象, 様子	

表 5: 獲得された課題・特長表現の例 [西山 10b]

課題表現	解決関係にあるとされた特長表現の例
信頼性…低下する	エロージョン…抑制する，信頼性…損なう…ない
ばらつき…生じる	エッチング…制御…容易だ，接触…確実だ-行う…できる，精度…大幅に - 向上する
歩留まり…低下する	歩留まり…高める…できる，歩留まり…低下…抑制する，歩留まり…向上する
不良…発生する	除去…容易に-できる，電流…低下…防止する，機械的 - 強度…向上する

援ツールのみであるが，このほかにもまだ様々な技術文書マイニングシステムが提案・実装されてよい．

差し当たっては，3.2 小節で紹介した，特定の技術課題を解決可能である技術の検索ツールの実装と評価が挙げられる．

5. おわりに

本稿では技術文書を対象とした情報編纂の一例として，特に特許文書から解決するべき技術課題，技術が提供する特長，および解決手段という三種類の意味を持った表現を抽出し，技術文書のマイニングに利用することを目指す一連の取り組みについて紹介した．三種類の表現の抽出方法については，抽出する表現に用いられやすい用言を中心とした構文パターンを利用して直接抽出する他にも，抽出する表現を伴いやすい文末表現を利用することで，分野に依存した表現を抽出することが可能であることや，系列ラベリング手法を利用してラベル付きデータから表現とその周辺の単語の特徴を学習し，抽出することが可能であることを示した．また，これらの表現の抽出結果を利用した技術文書マイニングの例として，新規な技術応用を発見するための技術調査支援ツールと，特定の課題を解決可能である技術の検索ツールを紹介した．

今後の課題としては，抽出した表現をより活用した技術情報アクセス手法の提案や，実際に技術情報アクセスが改善されることの評価が挙げられる．

参考文献

[Bagchi 00] Bagchi, S. and Tulskie, B.: e-business Models: Integrating Learning from Strategy Development Experiences and Empirical Research, in *20th Annual International Conference of the Strategic Management Society*, pp. 15–18 (2000)

[乾 06] 乾 孝司, 奥村 学：テキストを対象とした評価情報の分析に関する研究動向, 自然言語処理, Vol. 13, No. 3, pp. 201–242 (2006)

[加藤 06] 加藤 恒昭, 松下 光範：情報編纂 (Information Compilation) の基盤技術, 第 20 回人工知能学会全国大会予稿集, No. 1D3-2 (2006)

[Nanba 10] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-8 Workshop, in *Proceedings of the 8th NTCIR Workshop Meeting* (2010)

[Nishiyama 10] Nishiyama, R., Tsuboi, Y., Unno, Y., and Takeuchi, H.: Feature-Rich Information Extraction for the Technical Trend-Map Creation, in *Proceedings of the 8th NTCIR Workshop Meeting* (2010)

[Pang 08] Pang, B. and Lee, L.: *Opinion Mining and Sentiment Analysis*, Now Publishers (2008)

[Tjong Kim Sang 99] Tjong Kim Sang, E. F. and Veenstra, J.: Representing Text Chunks, in *Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, pp. 173–179 (1999)

[小林 05] 小林 のぞみ, 乾 健太郎, 松本 裕治, 立石 健二, 福島 俊一：意見抽出のための評価表現の収集, 自然言語処理, Vol. 12, No. 3, pp. 203–222 (2005)

[西山 09] 西山 莉紗, 竹内 広宜, 渡辺 日出雄, 那須川 哲哉：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, Vol. 24, No. 6, pp. 541–548 (2009)

[西山 10a] 西山 莉紗：特許公報を対象とした従来技術課題の抽出, 言語処理学会第 16 回年次大会, No. C1-3 (2010)

[西山 10b] 西山 莉紗, 竹内 広宜：同じ効果を持つ複数技術を同定するための知識抽出, 第 24 回人工知能学会全国大会予稿集, No. 2J2-NFC2-5 (2010)