3K3-2

技術文書マイニングのための特長表現抽出

Advantage Phrase Extraction for Mining Technical Documents

西山 莉紗 Risa Nishiyama　竹内 広宜 Hironori Takeuchi　渡辺 日出雄 Hideo Watanabe　那須川 哲哉 Tetsuya Nasukawa　武田 浩一 Kohichi Takeda

日本アイ・ビー・エム (株) 東京基礎研究所
Tokyo Research Laboratory, IBM Research

A novel semantic class of phrase, Advantage phrase, is defined in this paper. This class includes phrases such as “reduce cost,” “improve PC performance,” “provide early warning of a future failure” and other phrases mentioning a strong point of a new technology or product. These phrases often appear in patents, technology news articles, technology reports and other technical documents and very helpful for people who daily survey up-to-date technology to understand what differentiates the technology/product compared with others. This paper describes extraction method of the phrases and its moderate coverage for multiple typese of technical document: patents and release letters of an IT company.

1. はじめに

日夜様々な製品や技術が世に出されていく現代において，最先端の技術を常に捉え続けるのは容易なことではない．研究者やコンサルタントなど最先端の技術を把握する必要のある人々は，論文，特許，そして Web 上のテクノロジーニュースなどを日々調査しているが，新製品・新技術によってどのようなことが可能になるか，従来と比較してどのような長所を持っているか，ということをいち早く理解することは重要である．

本研究はこのような，大量の技術文書からある特定の技術エリアにおいて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を鑑みた，技術文書マイニング手法を提案する．**技術文書**とはここでは特許公報や科学技術論文，Web 上のテクノロジーニュース，メーカーや IT ベンダーなどのプレスリリースなど，主に技術概要や，技術によって実現される製品およびサービスについて記述している文書とする．技術文書の中には，新技術，新製品が持つ，好ましい性質や新機能について述べているものがある．例えば新たに発表された携帯電話について述べた文書では

「通話音質が向上する」
「着メロを高速にダウンロードできる」
「コンビニの買い物で小銭を出す必要がなくなる」

というような記述がなされている．このような表現を本稿では**特長表現**と呼ぶ．特長表現を技術文書から自動的に抽出することで，例えば最近の携帯電話市場で発表されている製品機能の動向や，最近の携帯電話技術を利用して実現できる事柄を，全ての製品概要や Web ニュースを読むことなく知ることができるだろう．このことは，最先端の製品，技術を短期間で理解するための技術文書マイニングにおいて，大変有用である．

本稿では以上に例示したような，製品や技術の特長を示す特長表現に着目し，これを技術文書から抽出する手法を与える．まず，特長表現の抽出対象である技術文書の種類と各々の特徴を説明し，特長表現抽出という新しいタスクと関連研究との位置づけを示す．次に特長表現の定義を述べ，その抽出法を説明する．最後に，公開特許公報と企業の新製品発表情報という，性質の異なる 2 種類の技術文書から人手と提案手法の両方を用いて特長表現を抽出し，人手による特長表現抽出の主観性や，提案手法の汎用性について検証した結果を示し，議論する．

連絡先: 西山 莉紗，日本アイ・ビー・エム (株) 東京基礎研究所，http://www.trl.ibm.com/people/lisa/，lisa@jp.ibm.com

2. 技術文書の種類と関連研究

2.1 技術文書の種類

技術文書は大まかに以下の 2 種類に分けることができる．

- 論文や特許，ホワイトペーパーなどの専門家向け文書
- Web ニュースや企業のプレスリリース，製品紹介などの一般向け文書

一般向け文書からの特長表現抽出は，専門家向け文書と比較してより困難なタスクになることが予想される．なぜなら専門家向け文書は話題の大半が技術説明に費やされるのに対し，一般向け文書は人事情報や経営状況など，技術と直接関係しない話題も含んでいることが多い．そのため，技術文書マイニングで注目される，技術に関連した特長以外の特長表現を抽出してしまう恐れがある．また，一般向け文書で用いられる表現は必ずしも記述的ではなく，体言止めや箇条書きなども多く用いられている．このような表現の多様性も，抽出タスクを難しくさせるだろう．

しかし，本研究で提案する特長表現抽出手法は，一般向け文書にも適用可能であることが望まれる．なぜなら一般向け文書は最新の技術情報について最新の顧客ニーズに合わせた用語を用いて記述しているのに対し，専門家向け文書は専門的な技術用語を用いて記述されていることが多い．例えば Web2.0 についての技術動向を知りたい場合，専門向け技術文書では Web2.0 という用語そのものが使われないことが多いため，「クライアント処理」「JavaScript」などの Web2.0 に関連する技術をあらかじめ洗い出しておき，それらの動向を調査する，という手順を踏む必要が生じる．

2.2 関連研究

筆者らは技術文書を対象として，ある技術によって可能となる事柄をリストアップする抽出しツールを提案した [西山 07]．その他にも，技術文書をマイニングし，企業の意思決定や新規ビジネス開拓の手がかりに利用しようという試みは既に多く行われ

ている [Losiewicz 00]．特に特許文書に対しては，特許文書の出願日や本文中に出現するキーワードなどを目的に応じた軸でまとめ，特許マップとして提示する技術 [新井 03, 市村 01, 渡部 03] が既に多く研究されており，特許マイニングツールとして広く利用されている．また，特許文書以外にも，競合他社が提供している意外な製品やサービスをマイニングによって発見し，意思決定に役立てようとすることを目的として，企業の Web サイト [Liu 01] や，科学技術論文 [Jacquenet 04] のマイニングが行われている．これらの既存研究は文書内に出現するキーワードに着目しており，特長表現というような，技術文書の中で重要な意味を持った表現に注目したものはこれまでに無かった．

一方，ある製品について情報収集するための情報抽出手法として，評価表現抽出 [乾 06] が広く用いられている．評価表現抽出は一般消費者がある製品について書いたレビューやブログ記事を対象として，そこから製品がどのように評価されているかを分析することを目的としている．言い換えれば，評価表現抽出は買い手である顧客のニーズを得るための技術であると言える．対して特長表現抽出は，売り手である企業が持っている技術シーズ (種) を得るための技術であると言える．

既存の評価表現抽出と本研究が提案する特長表現抽出との違いは，対象とするテキストデータだけでなく，抽出方法にもある．テキストから評価表現を抽出するためには，例えば「美味しい」という表現が好評を示し,「不味い」という表現が不評を示すというような，評価表現とその好評・不評を定義した評価表現辞書を予め何らかの方法で獲得しておき，これを利用する必要がある．対して特長表現抽出では，書き手である企業や技術者が考える新製品，新技術の好ましい特質の主張に重点を置いた文書を対象としている．そのため,「向上する」「可能にする」などの，長所を主張する際に用いられる用言パターンを利用することができる．このような用言の種類は，一般に評価表現辞書として獲得すべき表現の種類よりも少ないことが期待される．

3. 特長表現の定義

本研究では，特長表現を「技術によって実現される, 当該製品または技術の好ましい性質について述べた表現」と定義し，以下の 2 種類のクラスから成ると考える．

- Improve クラス: 元来その製品または技術が持っている好ましい点を伸ばして特長とする示唆する表現
- Reduce クラス: 元来その製品または技術が持っている望ましくない点を抑えて特長とすることを示唆する表現

この分類により，Improve クラスの表現は，その製品や技術にとっての好ましい要素が，反対に Reduce クラスの表現には好ましくない要素が述べられることになる．これらの要素を**特長対象**と定義する．

4. 特長表現の抽出方法

特長表現の抽出は図 1 の要領で行われる．まず文中に出現する，特長対象 (図中太字部) と，特長表現の手がかりとなる語 (図中下線部) の両方を同定する．その後本文を走査し，検出された手がかり語から前に戻る形でその前に出現する特長対象を抽出する．予め定められた深さ分戻って得られる表現を最終的な特長表現 (図中矩形部) とする．

本節ではまず手がかり語の詳細を示し，その後特長対象の同定方法を示す．

...**広告情報毎**に**広告**の**効果**などを迅速に**把握すること**ができる。

...**看板**を通して詳細**配布情報**を円滑に**入手すること**ができ、**看板**における**広告機能**を飛躍的に**向上させること**ができる。

図 1: 特長表現抽出の概要 (下線は同定された手がかり語，太字は特長対象として抽出された名詞句，矩形は特長表現として抽出される表現を示す)

4.1 抽出に利用する手がかり語

まず，表 1 に示すような用言ベースの手がかり語を用いて，特長表現の出現箇所を特定する．

表 1: 特長表現抽出に利用する手がかり語と抽出表現の例

手がかり語	抽出される表現例
Improve クラス	
～[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
Reduce クラス	
～[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

ここに示した手がかり語は限定的であり，各々のクラスを示す動詞は他にも数多く考えられるが，これだけで大半の特長表現を抽出できることが期待される．なぜなら，これらの手がかり語は主に特許文書に頻出する動詞から，前節で示したクラスのいずれかを示すと考えられるものを抽出し，作成したものであるためである．手がかり語は特許以外の技術文書に対しても汎用的に用いられるが，発明の，特に既存技術と比較した特長の記述に注力する特許文書から得られた頻出語は，特長表現の抽出に利用する上で高く信頼することができる．

4.2 特長対象抽出

手がかり語を利用して特長表現の出現箇所を特定した後，そこから予め定められた単語距離にある名詞，複合名詞ならびに名詞句を抽出する．手がかり語と同様に，文中に出現する全ての特徴対象は予め同定されている．本稿では手がかり語からさかのぼり，最も近くにある名詞句を特長対象として抽出する．

5. 評価実験

5.1 実験目的

本節では，2 種類の技術文書コーパスを利用して，以下の 2 点を検証する．

1. 特長表現抽出タスクの主観性 (subjectivity)
   正解データは技術文書から人手で技術特長を示している表現を抽出して作成する．このとき，2.2 小節で仮定したように，ある表現が製品や技術の特長を示すかどうかは書き手が用いる表現によってある程度客観的に判断でき

るものであり，過度に読み手 (被験者) の主観に依存しないものであることを検証する．

2. 複数種類の技術文書における，手がかり語の汎用性
2 節に示したように，技術文書には専門家向けの文書と一般向けの文書とがあり，それぞれ製品特長の表現方法も異なっていると考えられる．前節に示した手がかり語は主に特許文書から作られたものであるが，この技術者向け技術文書から作られた手がかり語がどの程度一般向け技術文書に対しても有用であるかを検証する．

5.2 利用する技術文書データ

本実験では専門家向けの技術文書の例として公開特許公報を，並びに一般向けの技術文書の例として新製品発表情報を利用する．

5.2.1 特許データ

技術文書リソースとして公開特許公報を利用する．本実験では 2006 年に公開された公開特許公報のうちランダムに抽出した 50 件を利用する．公開特許公報は定型文書であり，いくつかのセクションから成る．本実験で利用する公開特許公報データは XML で表記されているため，任意のセクションから特長表現を抽出することが可能である．本ツールでは「要約」中の「課題」「解決手段」セクション，そして「発明の開示」中の「発明の効果」セクションをこの後の処理に利用する．

5.2.2 製品発表データ

もう 1 つの技術文書リソースとして，企業の新製品・サービスの発表情報を利用する．本実験では日本アイ・ビー・エム株式会社で 2002 年から 2008 年の間に発表された製品・サービス発表レター[*1]50 件を利用した．この文書にはソフトウェア・ハードウェアの新製品情報だけではなく，新たなコンサルティングサービスの発表や，新価格設定の発表など，技術特長以外の情報も多く含まれている．本実験では，2 節に示した，一般向け技術文書が含む話題の多様性が技術特長の抽出精度を下げる，という仮説を検証するため，50 件の製品・サービス発表情報の中には，以下の 5 種類の内容のものを 10 件ずつ含めた．

- ソフトウェア新製品の発表
- ハードウェア新製品の発表
- 新サービスの発表
- 新プロモーション (販売企画) の発表
- 開発意向表明

なお，これらの区分はすでに各文書に定型情報として付けられていたものをそのまま利用した．

5.3 実験方法

ここではまず実験に使用した正解データの準備方法を述べ，そして準備した正解データと提案手法による特長表現抽出結果の分析方法について述べる．

5.3.1 正解データの準備

2 名の被験者に対して，特許データ 50 件と製品発表データ 50 件から，各々の文書で紹介されている技術および製品の特長を示しているとみなした表現を抜き出すことを依頼した．そして，それぞれの被験者から得られた特長表現を比較し，2 人の被験者間で一致した文字列を最終的な正解データとした．例えばある文書から被験者 A が「より小さな力で部品の取り外しを行うことができる」という表現を，被験者 B が「部品の取り外しを行うことができる」という表現を抽出した場合，2 つの抽出表現が重なり合っている「部品の取り外しを行うことができる」を正解の特長表現とする．なお，1 文書から抽出される特長表現は複数あり，1 文から抽出される特長表現も複数ある．

5.3.2 正解データと抽出結果の比較方法

- 特長表現抽出タスクの主観性
被験者間一致率 (inter-annotator agreement) を求めることで，タスクの主観性を評価する．被験者間一致率の指標としては kappa 統計量が広く用いられている．しかし，本タスクは定数試料の分類タスクではなく，1 文書，1 文から複数，任意長の表現を抽出するタスクであるため，kappa 統計量を利用することができない．kappa 統計量に代わる指標として式 (1) を用いて被験者間一致率を求める．

$$inter\text{-}annoatator\ agreement = \frac{2 \times n_{ab}(D)}{n_a(D) + n_b(D)} \quad (1)$$

ここで D は抽出対象となる文書集合，$n_a(D)$ は被験者 A が D から抽出した特長表現数，$n_b(D)$ は被験者 B の抽出表現数，そして $n_{ab}(D)$ は被験者 A と B の抽出表現の重なり部分から作られた，正解特長表現数を示す．

- 手がかり語の汎用性
手がかり語の汎用性は，各技術文書リソースについて precision(式 (2)), recall(式 (3)) および F 値 (式 (4)) を求め，これらの値によって評価する．

$$precision = \frac{TP(D)}{TP(D) + FP(D)} \quad (2)$$

$$recall = \frac{TP(D)}{TP(D) + FN(D)} \quad (3)$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \quad (4)$$

このとき $TP(D)$ は文書集合 D の正解特長表現と，提案手法が文書集合 D から抽出した表現が 1 文字以上重なった回数を示し，$FP(D)$ は提案手法が文書集合 D から抽出した表現のうち，正解特長表現のいずれとも重なり合わなかったものの数，$FN(D)$ は文書集合 D の正解特長表現のうち，提案手法が文書集合 D から抽出した表現のいずれとも重なり合わなかったものの数を示す．

5.4 実験結果と考察

人手による特長表現抽出の結果，特許データからは 259 表現，製品発表データからは 82 表現の正解データが得られた．ならびに，同じデータから提案手法によって特長表現を抽出した結果，特許データからは 227 表現，製品発表データからは 90 表現が抽出された．

ここでは本節の最初に挙げた，この実験で検証すべき事柄のそれぞれについて結果を示し，考察を行う．

5.4.1 特長表現抽出タスクの主観性

特許データ (Patent)，製品発表データ (Announcement Letter) それぞれについて被験者間の正解一致率を求めた結果を表 2 に示す．特許データにおける一致率が 80%を越えるのに対し，製品発表データにおける一致率はその半分近い 40%強に落ち込んでいることが分かる．

製品発表データにおいて被験者間で一致しなかった表現には，「特別価格で弊社製品をご購入いただける」「無償で新バー

*1 http://www-06.ibm.com/jp/domino02/NewAIS/aisextr.nsf/aissearch

ジョンのサポートを受けられる」というような，技術によって実現される特長ではなく，サービスや販売活動の特長を示している表現が見られた．このような表現は本研究が抽出を目的としている，技術特長を示す表現ではないため，事前に被験者に技術特長のみを抽出するよう指示，ないしトレーニングする必要があった．

また,「DVD R/W のサポート」「TCP/IP ソケット数の増加」など，技術特長ではあるが，特長を示していると判断するために，当該製品分野についての前提知識を必要とするものがあった．このような表現も特長表現として自動抽出されることが望まれるが，被験者の前提知識を必要とするため，高い信頼性が必要とされる正解データとして抽出されるべきものでは無いだろう．

表 2: 被験者間の正解一致率

Document set	Inter-annotator agreement
Patent	87.9%
Announcement Letter	43.5%

5.4.2 手がかり語の汎用性

特許データ (Patent)，製品発表データ (Announcement Letter) それぞれについて，抽出精度を求めた結果を表 3 に示す．ここでも特許データにおける F 値が 0.7 を上回るが，製品発表データでは 0.5 程度に落ち込むことが分かる．

表 3: 各技術文書リソースにおける抽出精度

Document set	Precision	Recall	F-measure
Patent	0.797	0.700	0.745
Announcement Letter	0.433	0.476	0.453

特許データと製品発表データ共に，誤検出となった特長表現の多くは 2 名のうちの一方の被験者によって，正解として抽出されていることが分かった．このことからは，前項で示した主観性が抽出精度に影響していることが分かる．

提案手法で抽出できなかった正解特長表現を見ると,「処理の高速化」「読影ワークフローの効率化」「低コスト」など，用言を含んでいない表現が見られた．このような表現は特に製品発表データに多く見られた．このような「効率化」や「高機能」,「短期間」といった特長表現は用言を含まない為，用言ベースの手がかり語を利用した提案手法では抽出できない恐れがある．しかし，このような体言のみからなる特長表現も,「効率化を可能にする」「高機能なシステムを実現する」といったように，表 1 に示した手がかり語を伴って記述される場合が多い．このような場合であれば特長表現として抽出可能であり，またこの抽出結果を利用して,「高機能」「短期間」のような特長を示す体言を自動獲得できることが期待される．このような，特長を示す体言への対応が今後の課題として挙げられる．

6. おわりに

本稿では製品や技術の好ましい性質や機能を示す，特長表現に着目し，これを技術文書から抽出する手法を提案した．まず，企業の意思決定のための技術文書マイニングの概要と，そこでの特長表現抽出の必要性について述べた．そして技術文書の種類とそれぞれの性質について説明し，特長表現抽出というタスクの関連研究との位置づけを示した．次いで，特長表現の定義とその抽出法を説明し，最後に，公開特許公報と企業の新製品発表情報という，性質の異なる 2 種類の技術文書から人手と提案手法の両方を用いて特長表現を抽出し，人手による特長表現抽出の主観性と，提案手法の汎用性という，2 つの点から結果を考察した．その結果，製品発表のように，特許と比較して幅広い話題の中で，幅広い表現を用いて特長が述べられるデータからであっても，比較的少ない手がかり語を利用して半数程度の製品・技術特長を抽出できることが分かった．この結果は，Web ニュースや企業のプレスリリースなどの幅広い文書からの特長表現抽出の可能性を示すものであり，技術文書マイニングを実用化する上で有効な結果であると言える．

今後の課題として，前章に述べた特長を示す体言の抽出の他に，分野全体を俯瞰できるようにするために，抽出された特長表現をまとめ上げる方法の検討が挙げられる．例えば「低コストを実現する」と「コストを削減する」とは，異なった手がかり語を持つが同等の内容を示している．このような同一の特長を示す表現をまとめ上げることで，ある製品分野において，どのような特長を持った製品がトレンドとなっているか，ということが分かる．その他様々なまとめ上げの手法を用いることで，より意思決定に有用な技術文書マイニングを実現できることが期待される．

謝辞 本稿で述べた技術文書マイニングや，そこで注目されるべき表現について，利用者としての立場から大変有用なコメントを頂戴した，アイ・ビー・エム ビジネスコンサルティングサービス (株) の前田潤治氏，笹本渡氏，倉持俊之氏に深く感謝いたします．

参考文献

[Jacquenet 04] Jacquenet, F. and Largeron, C.: Discovering Unexpected Information for Technology Watch, in *PKDD'04*, pp. 219–230 (2004)

[Liu 01] Liu, B., Ma, Y., and Yu, P. S.: Discovering unexpected information from your competitors' web sites, in *KDD'01*, pp. 144–153 (2001)

[Losiewicz 00] Losiewicz, P., Oard, D., and Kostoff, R.: Textual Data Mining to Support Science and Technology Management, *Journal of Intelligent Information Systems*, Vol. 15, No. 2, pp. 99–119 (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)

[市村 01] 市村 由美, 長谷川 隆明, 渡部 勇, 佐藤 光弘：テキストマイニング： 事例紹介 (<特集>「テキストマイニング」), 人工知能学会誌, Vol. 16, No. 2, pp. 192–200 (2001)

[新井 03] 新井 喜美雄：特許情報分析とパテントマップ, 情報の科学と技術, Vol. 53, No. 1, pp. 16–21 (2003)

[西山 07] 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)

[渡部 03] 渡部 勇：テキストマイニングの技術と応用 (<特集>情報の分析・解析法), 情報の科学と技術, Vol. 53, No. 1, pp. 28–33 (2003)