手がかり表現自動獲得による製品発表プレスリリースからの製品特徴の抽出

酒井 浩之　　**増山 繁**
sakai@tut.jp,　masuyama@tut.jp
豊橋技術科学大学 大学院工学研究科 情報・知能工学専攻

1　はじめに

現在，多くの製品の発表や発売に関するプレスリリースが，日々，発表されている．製品発表プレスリリースには，製品の概要と共に，その製品の特徴となる情報が含まれている．例えば，高画質を目指したテレビのプレスリリースには,「高画質を実現しました。」といった表現が出現している．製品発表プレスリリース中の，その製品の特徴に関する情報は製品のトレンド分析や企業の開発方針の決定等において重要な情報となる．しかしながら，製品発表プレスリリース中には製品特徴以外の情報も多く含まれている．そのため，本研究では，製品発表プレスリリースから製品特徴の情報を抽出する手法を提案する．

本研究では，例えば「高画質を実現しました。」,「大容量ＳＳＤを搭載。」といった，製品特徴の情報を含む文を製品特徴文と定義し，それを抽出する．その抽出には「実現しました。」といった手がかり表現を使用し，手がかり表現が含まれている文を製品特徴文として抽出する．しかしながら，製品発表プレスリリースから製品特徴文を抽出するために有効な手がかり表現は数多く存在し，そのような手がかり表現のリストを人手で作成することは手間がかかるうえ，網羅性の点からも容易ではない．例えば,「高画質」を製品特徴とする製品においても,「高画質を提供します。」「高画質を楽しむことができます。」といった表現が存在する．そのため，本研究では手がかり表現の自動獲得を行う．

手がかり表現は，我々が以前提案した交通事故事例記事からの事故原因表現の獲得のための手法 [5]，業績発表記事からの業績要因表現の獲得のための手法 [6] に基づいて行う．我々の既提案手法では，最初に少数の手がかり表現を人手で与え，手がかり表現をブートストラップ的に繰り返し獲得していく．しかしながら，再現率を上げるために多くの手がかり表現を獲得しようとすると，同時に不適切な手がかり表現も多く獲得されていき，それによって精度が大きく低下する [5]．そこで，不適切な手がかり表現を除去するための改良を施したうえで，本タスクに適用して取得する．

一般的に，ブートストラップ的な手法では，種としたインスタンス（本手法においては少数の手がかり表現に該当）と無関係なインスタンスが獲得されることがあることが知られている．Pantel らは，自己相互情報量に基づいた信頼度を導入して，Is-a のようなパターンとインスタンスをブートストラップ的に自動的に獲得する手法を提案している [4]．小町らは，Pantel らの手法に対して，グラフ理論に基づく分析を行い，無関係なインスタンスが獲得される現象である意味ドリフトが起こる原因を解析している [1]．本研究では，最初に少数の不適切な手がかり表現を人手で与え，不適切な手がかり表現をもブートストラップ的に繰り返し獲得することで，不適切な手がかり表現のリストを作成し，それらが手がかり表現として獲得されることを防ぐ．そのため，既提案手法 [5][6] に比べ，再現率を上げるために多くの手がかり表現を獲得しようとしても精度の低下を抑えることができる．

関連研究として，西山らは，公開特許公報と製品・サービス発表レターから「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」と定義される「特長表現」を抽出する手法を提案している [3]．西山らの手法では，人手で定めた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している．それに対して，本手法では少数の手がかり表現を入力することで，多くの手がかり表現を自動的に獲得することができる．また，製品発表プレスリリースには，例えば「画面タッチパネルで簡単操作」のように，体言止めで表される製品特徴文も多く存在するが，手がかり表現と同時に獲得する共通頻出表現（「簡単操作」など．詳細は後述）を使用することで，体言止めで表される製品特徴の抽出も可能とする．

2　提案手法

2.1　手がかり表現の自動獲得

手がかり表現自動獲得手法の概要を以下に示す（詳細は文献 [5][6] を参照.）.

手がかり表現自動獲得手法

Step 1: 少数の手がかり表現を人手で与える.

Step 2: 手がかり表現の直前に出現する可能性が高い表現を取得する.（以降，手がかり表現の直前に出現する可能性が高い表現を共通頻出表現と定義する．例えば「高画質」「省エネ」といった表現が抽出される．詳細は 2.2 節で述べる.）

Step 3: 共通頻出表現から，新たな手がかり表現を獲得する.（詳細は 2.3 節で述べる.）

Step 4: 獲得した手がかり表現から，新たな共通頻出表現を獲得する.（Step 2 と同一の処理）

Step 5: Step 2 から Step 4 を，新たな手がかり表現と共通頻出表現が獲得されなくなるか，もしくは，予め定めた回数まで繰り返す（図 1 を参照）. □

Step 1 では，初期の手がかり表現として「ができます。」,「を楽しめます。」,「が可能です。」を人手で与えた.

2.2　共通頻出表現の抽出

Step 2 において，手がかり表現の直前に出現する名詞を共通頻出表現の候補とする．例えば，手がかり表現「を楽しめます。」を使用して,「高画質を楽しめます。」という文からは「高画質」が共通頻出表現の候補となる．ただし,「こと」，および,「事」を共通頻出表現の候補から除外した．次に，手がかり表現から抽出された共通頻出表現候補の中から適切な共通頻出表現

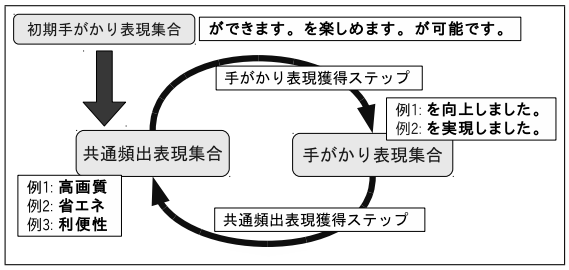

図 1: 手がかり表現自動獲得手法の概要

図 2: 共通頻出表現，手がかり表現の選別

を抽出する．具体的には，図 2 のように様々な手がかり表現に係っている共通頻出表現は適切であるという仮定に基づき，共通頻出表現が手がかり表現に係る確率に基づくエントロピーを求め，その値が閾値 T_e 以上の共通頻出表現を選別する．共通頻出表現が手がかり表現に係る確率に基づくエントロピーは式 1 で求める．

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s) \qquad (1)$$

ただし，手がかり表現を抽出する製品発表プレスリリース集合において，$P(e,s)$ は共通頻出表現 e が手がかり表現 s に係る確率，$S(e)$ は共通頻出表現 e が係る手がかり表現の集合である．閾値 T_e は，以下の式 2 によって設定する．

$$T_e = \log_2 |Ns| \qquad (2)$$

ただし，Ns は共通頻出表現を取得するのに使用した手がかり表現の集合，　は定数 (0 <　< 1) である．

2.3　共通頻出表現からの手がかり表現の獲得

共通頻出表現の抽出を行った後，抽出した共通頻出表現から新たな手がかり表現を獲得する．まず，抽出した共通頻出表現を含む文を抽出し，その中で共通頻出表現を含む文節 P_a が係っている文節 P_b を獲得する．次に，P_a に含まれる助詞を P_b に追加し，それを手がかり表現候補とする．ただし，文節 P_b に句点がある場合のみを手がかり表現候補とする．また，P_b が「しています。」「なります。」「しました。」「します。」であった場合，手がかり表現候補から除外する．例えば「省エネ」が共通頻出表現として取得されていた場合，「省エネを実現しました。」という文から「を実現しました。」が手がかり表現候補となる．また，手がかり表現候補に対しても，図 2 のように，様々な共通頻出表現が係っている手がかり表現は適切であるという仮定にもとづき，手がかり表現候補に対して共通頻出表現が係る確率に基づくエントロピーを式 3 で求め，閾値以上の候補を手がかり表現として抽出する．

$$H(s) = -\sum_{e \in E(s)} P(s,e) \log_2 P(s,e) \qquad (3)$$

表 1: 獲得された手がかり表現

を実現いたしました。 が向上します。 が出来ます。
が容易です。が特長です。 に優れています。
ができるようになりました。を搭載しています。

ただし，手がかり表現を抽出する製品発表プレスリリース集合において，$P(s,e)$ は手がかり表現 s に対して共通頻出表現 e が係る確率，$E(s)$ は手がかり表現 s に係る共通頻出表現の集合である．閾値は，共通頻出表現と同様に式 2 によって設定するが，Ns は新たな手がかり表現を獲得するのに使用した共通頻出表現の集合である．

2.4　不適切な手がかり表現の除去

上記の手がかり表現自動獲得手法では，手がかり表現獲得，共通頻出表現獲得の過程において不適切な手がかり表現を取得することがある．例えば，「を行います。」「を発売します。」といった表現が不適切な手がかり表現となる．「を行います。」が手がかり表現として抽出された場合，「保守サービスを行います。」といった文が製品特徴として抽出されてしまう．特に，定数　を小さい値に設定した場合，多くの不適切な手がかり表現が獲得され，大幅な精度の低下を招く．そのため，不適切な手がかり表現をストップワードとして保持し，手がかり表現の獲得の際に，それらの除去を行う必要がある．ここで，不適切な手がかり表現は，適切な手がかり表現以上に種類が多いため，それらをも自動的に獲得することを行う．不適切な手がかり表現の獲得手法は 2.1 節から 2.3 節で示した手法と同様であるが，初期手がかり表現として「を発売いたします。」「を発売します。」「を行います。」を使用する．そして，以下に示すように，手がかり表現獲得，不適切な手がかり表現獲得を交互に行うことで，多くの手がかり表現，および，不適切な手がかり表現を獲得する．

手がかり表現，不適切な手がかり表現の獲得手法

Step 1: 大きい値の定数　を設定して不適切な手がかり表現を獲得し，獲得された表現をストップワードリストに追加（大きい値の　を設定すると，式 2 で計算される閾値が高くなる.）.

Step 2: 定数　にて，手がかり表現の獲得を行う．その際に，ストップワードリストに含まれる表現を除去する．

Step 3: 定数　の値を小さく再設定し不適切な手がかり表現の獲得を行う．その際に，Step 2 で獲得された手がかり表現リストに含まれる表現を除去する．獲得された不適切な表現をストップワードとして，ストップワードリストを更新する．

Step 4: Step 2 と Step 3 を繰り返す. □

表 1 に，獲得された手がかり表現の一部を示す．我々の既提案手法では，定数　の値によって，共通頻出表現，および，手がかり表現獲得のために閾値が決定される．そして，　の値が大きいうちは，獲得される手がかり表現の数は少ないものの，精度が高い手がかり表現集合が獲得される．そのため，最初は大きい　の値を設

定して不適切な手がかり表現を獲得し，それらを除去して手がかり表現を獲得する．そして，　の値を小さく再設定したうえで，獲得された手がかり表現を除去して不適切な手がかり表現を獲得することで，適切な手がかり表現が不適切な手がかり表現として獲得されることを防ぐ．

2.5　文中に出現する手がかり表現の獲得

2.1 節の手法では，手がかり表現に句点がつくことを条件にしているため，文末に出現する手がかり表現しか獲得することができない．しかしながら，文末以外の文中（以降,「文中」とする）にも製品特徴を抽出するために有効な手がかり表現は多く存在している．例えば,「長時間駆動を実現したノートパソコン「○○」を商品化しました。」という文を製品特徴として抽出する場合,「実現した」という手がかり表現が必要である．(「を商品化しました。」という表現では製品特徴を含まない文も抽出される.）そこで，文中の手がかり表現を自動的に獲得する．

文中に出現する手がかり表現は，商品名等の特定の名詞を修飾している修飾節に出現していることが多い．そのような名詞を抽出し，それを修飾している文節から手がかり表現を獲得する．まず，名詞を修飾している文節で，かつ，文節に動詞を含んでいるものを獲得する．そして，以下の式 4 を用いて，ある名詞 n に対して，ある動詞を含む文節が修飾する確率に基づくエントロピー $H(n)$ を計算する．$H(n)$ が大きい名詞は，多くの種類の文節によって修飾されていることを表す．

$$H(n) = -\sum_{v \in V(n)} P(v,n)\log_2(P(v,n)) \qquad (4)$$

ただし，製品発表プレスリリース集合中において，$V(n)$ は，名詞 n を修飾する，動詞を含む文節の集合，$P(v,n)$ は，名詞 n が動詞を含む文節 v によって修飾される確率である．

ここで，名詞 n の重みを以下の式で定義する．

$$W(n) = 1/(1+H(n)) \qquad (5)$$

製品発表プレスリリース集合中で，動詞を含む文節によって一度しか修飾されない名詞の $H(n)$ は 0 になり，そのような名詞の重みは 1 となる．逆に,「こと」のような多くの文節によって修飾される名詞の $H(n)$ は大きくなるため，そのような名詞の重みは小さくなる．

次に，動詞を含む文節 v が，重み $W(n)$ が大きい名詞 n を修飾している確率に基づくエントロピー $H(v)$ を計算する．$H(v)$ が大きい値を割り当てられる文節 v は，製品発表プレスリリース集合中で動詞を含む文節によって一度しか修飾されないような，重み $W(n)$ が大きい名詞を多く修飾している文節となる．ここで，名詞の重み $W(n)$ が 0.7 以上の名詞を使用した．

$$H(v) = -\sum_{n \in N(v)} P(n,v)\log_2(P(n,v)) \qquad (6)$$

ただし，製品発表プレスリリース集合中において，$N(v)$ は，動詞を含む文節 v が修飾している名詞の集合，$P(n,v)$ は，動詞を含む文節 v が，重み $W(n)$ が大きい名詞 n を修飾している確率である．

表 2: 評価結果 (本手法)

	num_c	num_k	精度 (%)	再現率 (%)
0.7	19	13	81.9	46.0
0.6	28	76	81.1	51.1
0.5	68	160	80.4	55.5
0.4	186	212	78.4	62.9
0.3	434	351	72.2	67.8

ここで，$H(v)$ の値が大きい順に，上位 50 個の文節を手がかり表現として獲得した．獲得された手がかり表現には,「搭載した」「優れた」「対応した」といった表現があった．

2.6　製品特徴文の抽出

2.1 節で獲得された手がかり表現から追加した助詞を除去した表現が含まれている文（例えば,「を実現しました。」ならば「実現しました。」），または，2.5 節で獲得された手がかり表現を含む文を製品特徴文として抽出する．以降，2.1 節で獲得された手がかり表現を文末手がかり表現，2.5 節で獲得された手がかり表現を文中手がかり表現と定義する．また，製品発表プレスリリースには「高い操作性」といった，獲得した文末手がかり表現，文中手がかり表現が含まれない製品特徴文も多く存在する．このような文を抽出するために，本手法で獲得された共通頻出表現を使用する．共通頻出表現には,「操作性」や「信頼性」のように，製品特徴を表す語が含まれている．そこで，共通頻出表現を文末に含む文を製品特徴文として抽出する．

3　評価

本手法を実装し，評価を行った．実装にあたり，形態素解析器として ChaSen[1]，係り受け解析器として CaboCha[2][2]を使用した．製品発表プレスリリースとして，日経プレスリリース[3]から表題に「発表」を含む 15,995 件のプレスリリースを取得し，取得したプレスリリースから文末手がかり表現，文中手がかり表現，共通頻出表現を獲得した．共通頻出表現獲得，および，文末手がかり表現獲得の繰り返し回数を 5 回とした．正解データは，テストデータとして 100 件の製品発表プレスリリースを無作為に抽出し，人手で製品特徴文を抽出して作成した[4]．そして，獲得した文末手がかり表現，文中手がかり表現，共通頻出表現を使用してテストデータから製品特徴文を抽出し，正解データと比較することで，精度，再現率を求めた．評価結果を表 2 に示す．比較手法として，2.4 節で述べた不適切な手がかり表現の除去を行わない場合の評価結果を表 3 に示す．ここで，　は式 2 における，閾値を決定するための定数，num_c は獲得された手がかり表現の数，num_k は獲得された共通頻出表現の数である．また，表 4 に，文末手がかり表現のみ，文中手がかり表現のみ，共通頻出表現のみを，それぞれ使用した場合の精度，再現率を示す．なお，定数　の値を 0.4 とした．

[1]http://chasen.aist-nara.ac.jp/hiki/ChaSen/
[2]http://chasen.org/~taku/software/cabocha/
[3]http://release.nikkei.co.jp/
[4]テストデータの 100 件は手がかり表現を獲得するための製品発表プレスリリース集合から除外してある．

表 3: 評価結果 (不適切な手がかり表現の除去を行わない場合)

	num_c	num_k	精度 (%)	再現率 (%)
0.7	20	12	78.2	47.0
0.6	28	77	78.4	52.2
0.5	61	138	72.9	58.0
0.4	389	259	48.7	86.0
0.3	1168	484	47.6	89.7

表 4: 評価結果 (手がかり表現ごと)

	精度 (%)	再現率 (%)
文末手がかり表現のみ	86.1	45.3
文中手がかり表現のみ	72.5	19.3
共通頻出表現のみ	79.4	13.9
全て使用（本手法）	78.4	62.9

4 考察

表 2 と表 3 を比較すると，定数　が 0.4 の場合，本手法の精度は 78.4% であるのに対し，不適切な手がかり表現の除去を行っていない比較手法の場合は精度が 48.7% であり，定数　を小さくしていくと精度が大きく低下していく．定数　が 0.6 の場合までは，本手法，比較手法，ともに精度，再現率に大きな差がないことから，それより小さい値で不適切な手がかり表現が抽出されるようになることが分かる．本手法では，不適切な手がかり表現を自動的に獲得し，それを手がかり表現獲得時に除去することで不適切な手がかり表現が獲得されることを防ぐ．その結果，多くの手がかり表現を獲得し，再現率を上げるために定数　を小さく設定しても不適切な手がかり表現が獲得されず，精度が低下することを防いでいる．例えば，この処理を行っていない比較手法の場合では,「を目指します。」等の製品特徴文には出現しない手がかり表現が獲得されていた．その結果，比較手法では「さらなる売上拡大を目指します。」といった文が抽出されてしまった．それに対して，本手法では「を目指します。」は製品特徴文を抽出するための手がかり表現とは別に，不適切な手がかり表現として獲得されている．そのため，そのような文が製品特徴文として抽出されることを防ぐことができた．

しかしながら，不適切な手がかり表現として獲得された手がかり表現の中には，適切な手がかり表現も含まれていた．例えば「を図りました。」といった手がかり表現が不適切な手がかり表現として獲得された．その結果,「視認性の向上を図りました。」といった文が製品特徴文として抽出されなかった．一方，この手がかり表現は比較手法では獲得されている．しかしながら，この文には「視認性」や「向上」といった製品特徴を表す語が出現している．そして,「視認性」「向上」，ともに，共通頻出表現として獲得されているため，共通頻出表現を有効に使用することで，より再現率を高めることができると考える．

表 4 より，文末手がかり表現のみを使用した場合では再現率が 45.3% であるのに対し，文中手がかり表現，共通頻出表現を使用して製品特徴文を抽出することで，精度が 7.7% 低下したものの，再現率が 62.9% まで向上している．文中手がかり表現によって，例えば,「柔軟な接続性と利便性を兼ね備えたモデルです。」のように，文中手がかり表現「兼ね備えた」を使用することで，文末手がかり表現では抽出できない文を製品特徴文として抽出することができた．また，共通頻出表現によって，例えば,「操作もスイッチひとつの簡単操作」のような体言止めの文を共通頻出表現「簡単操作」を使用することで抽出することができた．しかしながら，共通頻出表現には製品特徴を表す語ばかりが獲得されている訳ではないため，精度の向上のために共通頻出表現から製品特徴を表す語を認定する必要がある．

5 まとめ

本研究では，製品発表プレスリリースから製品特徴文を，自動的に獲得した文末手がかり表現，文中手がかり表現を使用することで抽出した．また，文末手がかり表現を自動的に獲得する際に同時に獲得される共通頻出表現を使用することで，製品発表プレスリリースでは頻出する体言止めで終わる製品特徴文の抽出も可能とした．文末手がかり表現は我々の既提案手法 [5][6] を本タスクに適用して獲得するが，不適切な手がかり表現が獲得されることを防ぐために，不適切な手がかり表現をもブートストラップ的に繰り返し獲得することで，不適切な手がかり表現のリストを作成し，それらが手がかり表現として獲得されることを防ぐことができた．今後の課題として，精度の向上のために，共通頻出表現における製品特徴を表す語の認定を挙げる．

謝辞

本研究の一部は,（財）栢森情報科学振興財団研究助成金，科研費 若手研究 B(21700158)，総務省戦略的情報通信研究開発推進制度（SCOPE）地域 ICT 振興型の助成を受けたものである。

参考文献

[1] 小町守, 工藤拓, 新保仁, 松本裕治: Espresso 型ブートストラッピング法における意味ドリフトのグラフ理論に基づく分析ー語義曖昧性解消における評価ー, 人工知能学会論文誌, Vol. 25, No. 2, pp. 233–242 (2010).

[2] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842 (2002).

[3] 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).

[4] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, *Proceedinngs of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 113–120 (2006).

[5] 酒井浩之, 梅村祥之, 増山繁: 交通事故事例に含まれる事故原因表現の新聞記事からの抽出, 自然言語処理, Vol. 13, No. 4, pp. 99–124 (2006).

[6] Sakai, H. and Masuyama, S.: Cause Information Extraction from Financial Articles Concerning Business Performance, *IEICE Trans. Information and Systems*, Vol. E91-D, No. 4, pp. 959–968 (2008).