特許公報を対象とした従来技術課題の抽出

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

1　はじめに

近年，特に特許公報を対象として，「コストを削減することができる」や「小型化が可能になる」といった，個々の技術で可能になることや，それがもたらす効果を表す表現を抽出し，特許分析の補助を行う試みがある [5, 7]．これらの表現は同時に「コストが高い」,「装置が大きくなる」というような，その領域における従来技術では解かれていなかった課題（**従来技術課題**）を解決していることを示唆している．

科学技術論文や特許公報などの技術文書では，しばしば当該技術の効果と，それによって解決される従来技術課題とが同じ文書内に記述されている．そのため，技術文書から技術の効果だけでなく，従来技術課題も抽出することは，技術の長所を理解する上で相補的な役割を果たすことが期待される．

以上より，本研究ではまず「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を**課題表現**と定義し[1]，これを技術文書から自動抽出することを試みる．

抽出対象の技術文書として，特許公報の「発明が解決しようとしている課題」セクションを利用する.「発明が解決しようとしている課題」は従来技術課題を中心に記しているセクションだが，図 1 の例の後半にあるように，効果も併せて記述されている．上記セクションから図中下線部にあるような従来技術課題を抽出することは，科学技術論文の要約を始めとした，他の種類の技術文書からの抽出にも役立つことが期待される。

本稿ではまず，関連する既存の研究について述べ，そして，課題表現の抽出方法について説明した後，評価実験の結果を紹介する．

2　関連研究

課題表現は従来技術に対する悪い評価を表していると言えるため，評価表現抽出で用いられてきた辞書を用いた抽出アプローチ [2] を用いて，不評の表現を検出することにより同定できると考えられる．しかし，これまでに中心的に扱われてきた商品のレビューテキストやブログテキスト中の評価表現と比較して，広いドメインの表現を抽出する必要があるため，既存の評価表現辞書をそのまま利用することは難しい．

【発明が解決しようとする課題】
ところが、従来技術では、軸方向空隙形の超小形モータの場合、フレームの段付部からロータの端面までの長さの精度を保つには、特に軸受、軸受ハウジング、ロータなどの各部の軸方向寸法の機械加工精度を高く維持する必要があり、加工時間、加工コストの増加を招くという問題があった。さらに、コギングトルクを少なくするため、ステータにスロットを設けずに電機子コイルをステータの表面に接着などにより固定した場合、空隙の長さは電機子コイルの表面からロータの表面までの長さとなるが、電機子コイルを機械加工することができないので、空隙長さの精度を維持することが難しく、空隙長さを大きめに決めることになり、発生トルクが低下するという問題があった。本発明は、電機子コイルの精度に関係なく、空隙長さを可能な限り小さく調整できるようにすることを目的とするものである。

図 1:「発明が解決しようとしている課題」の記述例

本研究が抽出する課題表現は，従来の評価表現抽出の中で「客観的な評価表現」[4] とされていた種類の表現であると言える．[4] では，例えば「がん」や「健康」などの，一般に好ましくない意味または好ましい意味で用いられやすい名詞を収集し，それぞれに不評と好評の極性を振った辞書を利用することで，客観的な評価表現の抽出を可能にしていた．しかし，課題表現には様々な専門用語が含まれており，それらの全てに好評・不評の極性を振った名詞辞書を作成することは困難である．また，[1] では，製品の長所と短所を分けて記したレビューテキストを利用し，例えば「long battery life」という表現が長所のセクションに出現しやすいことを利用して,「long battery life」という客観的な表現の極性を獲得していた．これに対し，本研究の手法は長所と短所が明確に分離されていないテキストから客観的な評価表現を抽出している点が異なる．

一方，要望表現抽出 [3] や効果表現抽出 [7] などでは，「〜が欲しい」や「〜できる」というような，書き手の要望や技術の効果を表す文脈表現を活用した抽出手法が提案されている．本研究は，これらの研究で用いられていた手法を，課題表現抽出という新しいタスクに適用することの有用性を検証するものである．

[1]なお，酒井ら [7] が「技術課題情報」と定義している表現は，本文中で示した「小型化が可能になる」のような効果を示す表現であり，本研究で抽出を試みる「従来課題」とは異なる．

3 課題表現の抽出

3.1 アプローチ

レビューテキスト中の評価表現と同様に，課題表現は「半導体製造装置のコストが上がる」や「歩留まり+が+高い」のように,「名詞句+助詞+動詞」または「名詞句+助詞+形容詞」の構造をなしていることが多い．従って本研究では，図 2 の粗点線部にあるように，単語の依存構造木の中から 1 節で定義した意味を持つ「名詞句-助詞-動詞」または「名詞句-助詞-形容詞」の部分木を課題表現として抽出する．なお，最後の動詞および形容詞が助動詞を修飾している場合は，それも併せて抽出する．従って,「情報量を増大する」と「情報量を増大しない」（情報量を増大する+ない）は区別される．

課題表現は分野固有の表現を多く含む．そのため，例えば「アーチファクトが発生する」という表現が X 線検査装置の技術課題を示すということは，分野外の人間には分かりにくい．しかし，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．このような文末表現は，記述内容が当該技術領域で望ましくない事柄であり，解決されるべき課題であることを示唆する文脈を形成する．本研究ではこの文末表現を**課題文脈パターン**と呼び，課題表現の抽出に利用する．

また，課題文脈パターンを伴わずに出現する課題表現ももちろん多く存在する．そこで，課題文脈パターンを修飾する形で出現しやすい課題表現（図 2 細点線部参照）を利用して**課題表現辞書**を自動獲得し，課題文脈パターンと併せて利用する．

3.2 抽出手順

本研究では，2 節で述べた要望表現の抽出 [3] の手法を参考にし，図 3 の手順で課題表現を抽出する．

まず，出現頻度が高く，かつ信頼性が高い課題文脈パターン集合 P をシステムに与える．

次に，大量の文書を解析して課題文脈パターン $p \in P$ に係りやすい課題表現辞書エントリ候補を獲得する．課題表現辞書エントリ候補は，依存木上で課題文脈パターン以下に出現する課題表現のうち，名詞句を主辞（すなわち，依存木上で頂点にある名詞）に代表させたものである（図 3 中 2 の実線部参照）．これは，頻出表現を獲得するにあたり，表現のバリエーションを少なくするためである．そして，課題文脈パターン集合 P に係っている全ての課題表現辞書エントリ候補 $i \in I'$ に対して信頼度 r_i（式 (1)）を求める．信頼度上位 k 件の課題表現を選択し，課題表現辞書 $I \subseteq I'$ とする．

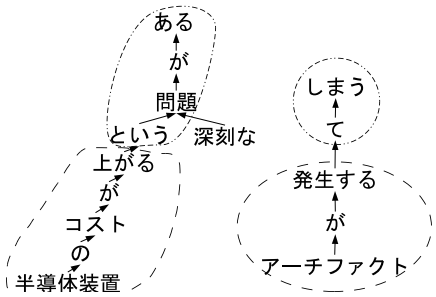

図 2: 課題表現（粗点線部）および課題文脈パターン（細点線部）の例

$$r_i = \frac{\sum_{p \in P} \mathrm{freq}(i,p)}{\mathrm{freq}(i)} \tag{1}$$

ここで，$\mathrm{freq}(i,p)$ は課題表現辞書エントリ候補 i が課題文脈パターン p に係っている頻度を示し，$\mathrm{freq}(i)$ は課題表現辞書エントリ候補 i の出現頻度を示す．

課題表現辞書は特定の分野や装置に依らない技術課題を示している表現を収集する目的で作成する．しかし，依然として，特定の分野や装置のみで扱われている技術課題も存在する．そこで，課題表現抽出の際には，課題表現辞書を用いて，課題文脈外に出現する課題表現を抽出するとともに，課題文脈パターンを利用して，特定の分野や装置固有の課題表現も抽出する（図 3 中 3 の粗点線部参照）．

4 評価実験

本節では，最初にシステムに与える課題文脈パターンのみを利用した課題表現抽出結果と，前節で提案した手法によって獲得した課題表現辞書を併せて利用した抽出結果との間で精度を比較する．

課題表現辞書獲得用の特許公報データとして，2003 年から 2006 年までの間に特許として成立した特許公報のうち,「発明が解決しようとしている課題」セクションを有している公報 127,028 件を利用した．

上記データから提案手法により課題表現辞書を自動構築した後，学習に利用したデータとは別の特許公報 2,897 件から，課題表現を抽出した．

実験で用いた課題文脈パターンを表 1 に示す．これは，学習データ中に出現する文末表現のうち，出現頻度上位 100 位以内のものから，人手で課題文脈を示すものを選択し，作成したものである．

そして，表 1 の課題文脈パターンを利用して得られた課題表現辞書エントリ候補から，出現頻度が 6 件以上のものを絞り込んだ結果，16,119 件の候補が獲得された．獲得された課題表現辞書エントリ候補から，上位 10%(1.611 件)，20%(3,222 件)，30%(4,833 件)，

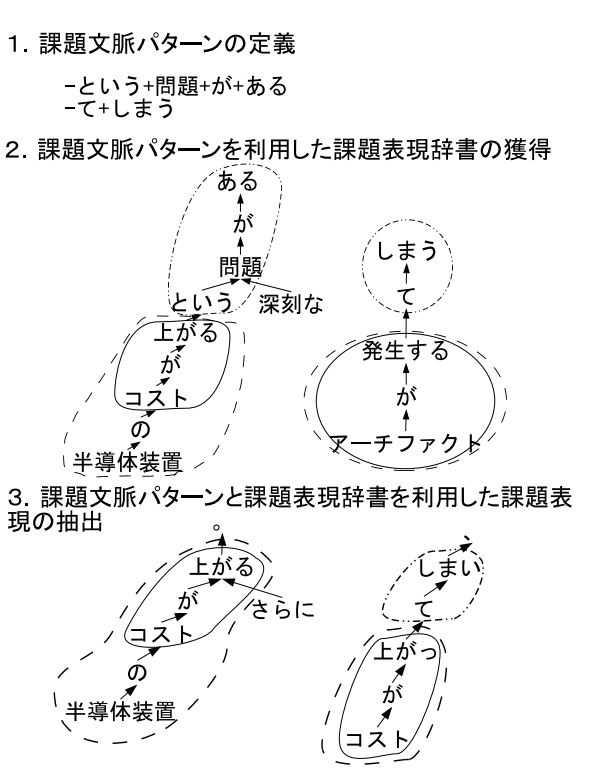

図 3: 課題表現の抽出手順

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，

表 1: 実験で使用した課題文脈パターン（カッコ内はふりがなを表す）

40%(6,444 件), 50%(8,055 件) を選択して課題表現辞書を作成し，課題表現抽出に利用した．

テストデータ 2,897 件中に含まれる全ての課題表現を人手でアノテーションすることは困難である．また，本実験では少数の限定された分野の文書からの抽出精度ではなく，様々な課題表現を含む広い分野の文書からの抽出精度を比較したい．そのため，本実験ではそれぞれの手法及び課題表現辞書作成方法によって抽出された課題表現からランダムに 100 件ずつ取り出し，技術課題を表しているものを正解として，意味の通らないもの，不完全な表現が抽出されているものについては不正解としてそれぞれ適合率を求めた．再現率の比較に当たっては [6] で定義されている相対再現率 $R_{A|B}$ (式 (2)) を用いて，課題文脈パターンと課題表現辞書を利用した場合の再現率 (R_A) と課題文脈パターンのみを利用した場合の再現率 (R_B) の比較を行った．

$$R_{A|B} = \frac{R_A}{R_B} = \frac{C_A/C}{C_B/C} = \frac{C_A}{C_B} = \frac{P_A \times |A|}{P_B \times |B|} \qquad (2)$$

ここで，C_A は課題文脈パターンと課題表現辞書を併用した場合の正解抽出数，C_B は課題文脈パターンのみを利用した場合の正解抽出数，C はテストデータ中の全正解数を示す．また，P_A，P_B はそれぞれ課題文脈パターンと課題表現辞書を併用した場合の適合率と，課題文脈パターンのみを利用した場合の適合率，そして $|A|$，$|B|$ は課題文脈パターンと課題表現辞書を併用した場合の全抽出表現数，課題文脈パターンのみを利用した場合の全抽出表現数を示す．

抽出表現数と適合率および相対再現率を表 2 に示す．課題文脈パターンのみを利用した抽出結果 (Base pattern) では，課題文脈中に現れる課題表現のみ抽出されるため，理論上は適合率が 1 となる．しかし，実際は，課題表現の構造を「名詞句+助詞+動詞または形容詞」に限定していることによる，主要な情報を取りこぼしや，係り受け解析の誤りなどにより，課題として意味をなさない無効な表現が入ってしまっているため，適合率は 1 にならなかった．意味をなさない課題表現の抽出を避けるためには，抽出する課題表現の構造に工夫が必要となる．

同じく表 2 より，提案手法によって獲得された課題表現辞書を利用すると，相対再現率が向上し，課題文脈上に無い課題表現も広く抽出できていることが分かる．獲得された課題表現辞書エントリの例を表 3 に示す．「パーティクル+が+増加する」「アーチファクト+が+発生する」など，分野特有の課題表現が獲得されていることが分かる．

その一方で，「情報+が+消える」，「針+が+曲がる」などの，分野や装置の種類によっては必ずしも解決するべき課題とはならないと考えられる表現も獲得されている．このようなエントリの混入は信頼度の計算に分野依存性の情報を導入することで軽減されることが期待される．例えば今回のように特許公報から課題表現辞書を作成する場合には，特許に付与されている IPC コードの利用が考えられる．またその他の種類の技術文書でも，文書の主題となっている技術の名称などを利用することである程度分野依存性を考慮できることが期待される．

評価実験の結果全体を通し，課題表現辞書エントリ候補の上位 50%を抽出に利用しても，適合率を 7 割程度に保ったまま，課題文脈パターンのみを利用した場合のおよそ 2 倍の再現率で正解エントリを抽出できている見込みがあることが分かった．このことから，提案手法によって課題表現を特許公報の「発明が解決しようとする課題」セクションから高い精度で抽出できていることが分かる．

5 おわりに

本稿では課題表現という，技術課題を示す客観的な不評表現に注目し，課題文脈を利用した課題表現辞書の自動獲得手法と，それを利用した課題表現の自動抽出

抽出方法	抽出数	適合率	相対再現率
Base patterns	2505	0.79 (79/100)	1
+Top 10%	2771	0.77 (77/100)	1.08
+Top 20%	3504	0.82 (82/100)	1.45
+Top 30%	4289	0.69 (69/100)	1.50
+Top 40%	4856	0.71 (71/100)	1.74
+Top 50%	5620	0.71 (71/100)	2.02

表 2: 課題表現抽出結果

情報+が+消える，誤り+に+弱い，
熱+により+劣化さ+せる，針+が+曲がる，
マイクロプロセッサ+が+占有さ+れる，
容器+から+溢れる，アーチファクト+が+発生する，
ディスク装置+が+大型化する，
コイル+が+焼損する，パーティクル+が+増加する

表 3: 獲得された評価表現辞書エントリの例（信頼度上位 10%から抜粋）

可能性を示した．今回の報告は特許公報の「発明が解決しようとする課題」セクションを対象としたものであったが，他の種類の技術文書に対しても，本手法を適用できると考えられる．今後の課題として，科学技術論文要旨などの他の種類の技術文書における提案手法の有効性の検証と，抽出された課題表現を利用した技術文書分析の有効性の検証が挙げられる．

参考文献

[1] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pp. 241–248, 2008.

[2] 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.

[3] Hiroshi Kanayama and Tetsuya Nasukawa. Textual demand analysis: Detection of users' wants and needs from opinions. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pp. 409–416, 2008.

[4] 中川哲治, 宮森恒, 赤峯享, 乾健太郎, 黒橋禎夫. Web 上の客観的記述からの評価情報抽出に関する技術的検討. 言語処理学会第 14 回年次大会発表論文集, 2008.

[5] 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.

[6] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING/ACL '06)*, pp. 113–120, 2006.

[7] 酒井浩之, 野中尋史, 増山繁. 特許明細書からの技術課題情報の抽出. 人工知能学会論文誌, Vol. 24, No. 6, pp. 531–540, 2009.