2J2-NFC2-5

同じ効果を持つ複数技術を同定するための知識抽出

Extracting Problem-Advantage Pairs from Patent Disclosures

西山 莉紗　　竹内 広宜
Nishiyama, Risa　　Takeuchi, Hironori

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

Grouping technologies having the same or similar effects is helpful for finding technical trends in the field or finding a solution of a specific business issue. This work aims at finding those technologies from patent disclosures by extracting pairs of problem phrase and advantage phrase in a "solving" relationship. By utilizing the sectioning in patent disclosures, the problem-advantage pairs are extracted to be utilized in the patent search. Results of an experiment utilizing Japanese patent disclosures showed that the proposed method is able to find problem-advantage pairs that are difficult to find the relationship between them from their word similarities.

1. はじめに

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．特許公報や科学技術論文，ならびにホワイトペーパーなどの技術文書は，新技術の調査に役立つ情報源であり，ここから特定の効果を持つ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．

このとき，ある技術が持つ効果は様々な表現で表される．例えば図 1 に示すように，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．しかし同時に,「ユーザーに複雑な操作を要求しない」技術も間接的に操作性の問題を解決していると言え,「作業効率を大幅に向上する」技術も，特にユーザーからのインタラクティブな操作を必要とするようなシステムでは操作性の問題の解決につながることが期待できる．

本研究では，上記に示したような，同じ課題の解決につながる，すなわち同じ効果を持っている複数技術を同定することを目的として，特許公報から技術課題を表す表現と，それを解決または和らげることが期待される特長表現を抽出するタスクを提案する．4. 節で説明するように，特許公報では,「発明が解決すべき課題」セクションに従来技術が抱えていた技術課題が記述されており,「発明の効果」セクションに提案技術の特長が記述されている．本研究ではこれらの２セクションを利用して，課題とそれを解決する特長表現とを抽出する方法を提案する．

本稿ではまず，解決関係にある課題・特長表現抽出タスクを定義し，タスクに関連する従来研究を紹介する．次に，抽出に利用する特許公報の構成を示し，抽出に利用する「発明が解決すべき課題」セクション，ならびに「発明の効果」セクションに記述されている事柄を説明する．そして，課題・特長表現対抽出の全体的なシステム概要について説明し，上記２セクションを利用した抽出方法について述べる．最後にＩＣチップや半導体製造装置が属する技術分野を対象とした，抽出方法の検証実験について述べる．

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/

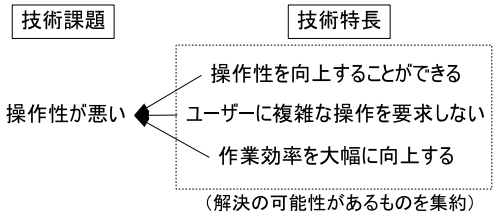

図 1: 同じ効果を持つ技術の集約イメージ（情報処理システムの例）

2. 解決関係にある課題・特長表現対抽出タスクの定義

本タスクは以下のように定義できる．まず，既存研究 [西山 10] と同様に，当該技術分野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呼ぶ．並びに，当該技術の新たな長所を示している表現を特長表現 [西山 09] と呼ぶ．

課題表現 i と特長表現 a が解決関係にあるとき，特長表現 a で表される長所を持った技術は，技術領域 f において，課題表現 i で表される技術課題を解決または和らげることが可能である．

3. 関連研究

本研究が目指す，同じ効果を示す複数の表層的に異なる表現の集約は，自然言語処理分野における言い換え (paraphrasing) 研究と関連が深い．従来の言い換え表現抽出タスクでは，２ヶ国語 [Bannard 05] または１ヶ国語 [Barzilay 01] で同じ内容を記したパラレルコーパスを用いた手法が主流であった．しかし，著者らが知る限り，本研究が扱う解決関係にある文を対にしたパラレルコーパスは存在しないため，これらの手法をそのまま適用することはできない．

２つの文が含意または矛盾関係にあることを判定する，含意関係認識（Textual entailment recognition）研究も，本研究と関係が深いと言える．このタスクでは，例えば「Mack Sennett was involved in the production of "The Extra Girl".」という文に対し、「"The Extra Girl" was produced by Sennett.」

という仮説を照合して，両者間の含意・矛盾関係を判定することが目的である．これに対し，本研究は文書中の課題表現と特長表現との間に,「解決関係」という，これまでの含意研究で扱われてこなかった新しい関係を見出すことを目的としている．含意・矛盾関係と異なり，解決関係の推定では，従来手法（[Hickl 06, MacCartney 08] など）で用いられてきた WordNet などの言語資源上の類義反義関係を直接利用することができない．また，従来の含意研究の多くが文同士の関係推定を扱っていたのに対し（例えば Pascal Recognizing Textual Entailment (RTE) Challenges [Dagan 05] における取り組みなど），本研究では，課題を表す短い表現と，文中から抽出された特長を示す表現との関係を推定する点も異なる．

技術文書からの情報抽出という観点では，NTCIR-8 の特許マイニングタスクのサブタスクとして，特許公報と科学技術論文から技術特長を表す表現を抽出する試みが行われている [Nanba 10]．本タスクによる課題との解決関係に基づく特長表現の集約は，このサブタスクが最終的に目指している技術動向マップの自動作成にも役立つことが期待される．また，特許公報からの情報抽出として，坂地らが「装置を小型化することにより，省スペース化を実現することができる.」というように,「ことにより」などの論理的接続を示す表現を利用して「装置を小型化する」という，技術が直接的に提供する効果と,「省スペース化を実現する」という，ユーザーに享受できる便益を示す効果との対を抽出している [坂地 09]．この結果を利用して，例えば「装置を小型化する」と「省スペース化を実現する」が同じ効果をもたらすという関係を得ることも可能である．しかし，坂地らの手法が「ことにより」のような接続指標でつながれて記述されている表現を対象としているのに対し，本手法では同一文書に出現するが，必ずしも接続指標で連結されているとは限らない表現も集約可能である点が異なる．また，効果と対になる課題表現を扱い，効果表現との関係を推定している点も異なる．

4. 特許明細書の構成

特許公報には請求項の他に,「発明の詳細な説明」として，その発明が属する技術分野や，先行技術文献を書く項目を設ける必要がある．その中の「発明の概要」という章の中に，以下の３つの節を設けることが推奨されている．

- 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

「発明が解決しようとする課題」節は，科学技術論文の序論（introduction）に相当する節である．ここでは，発明と関連が深い従来技術の問題点が中心に記述される．また，公報によっては，科学技術論文と同様に，最後に発明が持つ長所や，期待できる効果を簡単に記している場合もある．

対して,「発明の効果」節は，科学技術論文の結び（conclusion）に相当する節である．ここでは，その前の「課題を解決するための手段」で書かれた発明の構成が,「発明が解決しようとする課題」に書かれた従来技術と比較してどのように優れているか，ということを説明する．

以上のように，一般に「発明が解決しようとする課題」では発明が乗り越えている従来技術の課題が書かれており（以降，このセクションを**課題セクション**と呼ぶ),「発明の効果」では発明がどのように課題を乗り越えているか，ということについて書かれていることが期待される（以降，このセクションを**効果セクション**と呼ぶ）．次章では，この２セクションを利用した課題・特長表現対の抽出方法について説明する．

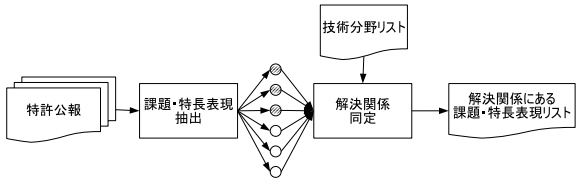

図 2: システム概要

5. システム概要

図 2 にシステム概要を示す．まず，特許公報の課題・効果セクションから課題表現と特長表現を抽出する．そして，6. 節で説明する解決関係同定手法により，抽出された課題表現と特長表現のうち，出現文書頻度が 5 以上のものの全ての組み合わせについて，解決関係が成り立つかどうかを評価する．

ある課題表現がある特長表現と解決関係にあるか否かは，技術調査を行いたい分野によっても異なる．そのため，解決関係の同定は技術分野ごとに行う．

公報中で，発明が関連する技術分野は複数の国際特許分類コード（IPC コード）によって示される．IPC コードは木構造の分類を示しており，上位分類からセクション（A から H までの 8 種類），クラス（セクションによって 5〜36 種類），サブクラス，メイングループ，サブグループとなっている．本研究では，各公報に割り当てられている IPC コードを用い，その発明が属する技術分野とする．

6. 課題・特長表現対の抽出

まず，特許公報の課題・効果セクションから課題表現 $\boldsymbol{p}$ と特長表現 $\boldsymbol{a}$ を抽出する．まず，課題・効果セクションを形態素解析，および構文解析し，両セクション中に出現する単語の係り受け構造（依存構造）を得る．そして，課題表現 i と特長表現 a として，課題・効果セクションからそれぞれ図 3 に示すテンプレートに合致した依存構造を持つ表現を抽出する．

課題表現 $i \in \boldsymbol{p}$ に対する，特長表現 $a \in \boldsymbol{a}$ の解決関係になりやすさ $s(i,a)$ は，相互情報量を用いて算出する．i と a との相互情報量は，課題表現 i が課題セクション中に出現する文書頻度 $\mathrm{DF}_p(i)$，特長表現 a が効果セクション中に出現する文書頻度 $\mathrm{DF}_e(a)$，課題表現 i と特長表現 a とが，同じ公報中で共起する文書頻度 $\mathrm{DF}(i,a)$ を用いて，以下のように算出できる．

$$\mathrm{pmi}(i,a) = \log \frac{p(i,a)}{p(i)p(a)} \propto \frac{\mathrm{DF}(i,a)}{\mathrm{DF}_p(i)\mathrm{DF}_e(a)}$$

しかし，上記の式によって算出した相互情報量は低頻度の表現に対して特に大きい値を返すことが知られている [Pantel 06]．そのため，本研究では [Pantel 06] による改良を用い，以下の式で相互情報量を算出する．

$$\begin{aligned}&\mathrm{pmi2}(i,a)\\ &= \log \frac{p(i,a)}{p(i)p(a)} \times \frac{\mathrm{DF}(i,a)}{\mathrm{DF}(i,a)+1} \times \frac{\min(\mathrm{DF}_p(i),\mathrm{DF}_e(a))}{\min(\mathrm{DF}_p(i),\mathrm{DF}_e(a))+1}\\ &\propto \frac{\mathrm{DF}(i,a)}{\mathrm{DF}_p(i)\mathrm{DF}_e(a)} \times \frac{\mathrm{DF}(i,a)}{\mathrm{DF}(i,a)+1} \times \frac{\min(\mathrm{DF}_p(i),\mathrm{DF}_e(a))}{\min(\mathrm{DF}_p(i),\mathrm{DF}_e(a))+1}\end{aligned}$$

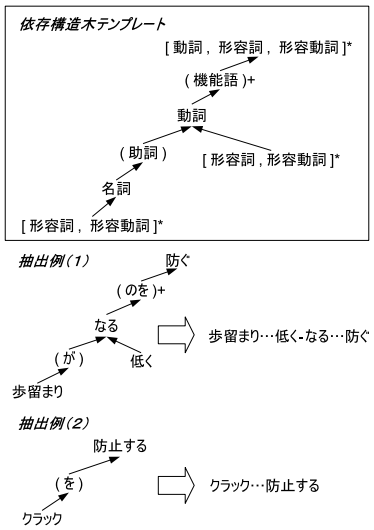

図 3: 課題・特長表現として抽出する依存構造木テンプレートと抽出される表現の例．テンプレート中の角括弧で囲まれたノード（単語）は，角括弧内のいずれかの品詞の単語にマッチし，丸括弧で囲まれたノードは，表現中で「…」という文字列に集約される．また，“+” “*”が付いたノードは必須ではなく，該当するノードがあった場合に表現に付与し，“+”が付いたノードは二つ以上存在する場合にはまとめて省略される．

上記の補正済み相互情報量を用い，分野 f における課題表現 i と特長表現 a の解決関係になりやすさを示す評価指標 $s(i,a|f)$ を以下のように算出する．ここで，$\mathrm{DF}_f(i,a)$ は，分野 f に属する特許公報における，i と a の共起文書頻度，$\mathrm{DF}_{fp}(i)$，DF_{fe} はそれぞれ，分野 f に属する特許公報の課題セッションに課題表現 i が出現する文書頻度，分野 f に属する特許公報の効果セッションに特長表現 a が出現する文書頻度を示す．

$$
\begin{aligned}
&s(i,a|f)\\
&= \frac{\mathrm{DF}_f(i,a)}{\mathrm{DF}_{fp}(i)\mathrm{DF}_{fe}(a)} \times \frac{\mathrm{DF}_f(i,a)}{\mathrm{DF}_f(i,a)+1}\\
&\times \frac{\min(\mathrm{DF}_{fp}(i),\mathrm{DF}_{fe}(a))}{\min(\mathrm{DF}_{fp}(i),\mathrm{DF}_{fe}(a))+1}
\end{aligned}
$$

この評価値が一定値以上となる i，a を抽出することで，解決関係にある課題・特長表現対の抽出とする．

しかし，上記の評価指標は課題セクションと効果セクションとの間の表現の共起を利用しているため，例えば同一発明者による同一分野への複数の出願など，記述が類似した公報がある場合，解決関係にない表現も共起しやすくなり，誤って抽出されてしまう可能性が高い．

このようなノイズとなる表現対に低い評価値を与えるため，本研究ではブートストラップ的アプローチを用いる．評価値を算出する際に，分野 f の文書集合からある一定のサイズ n のサブセットを取り出し，それを M 回繰り返して評価値を算出する．評価値の算出手順は以下のようになる．

- $k = 1, 2, \ldots m$ について
  - 分野 f の文書集合 D_f からランダムに n 件の文書を選択する (D_{fk})．
  - 文書サブセット D_{fk} を用いて，評価値 $\hat{s}_k(i,a|f)$ を算出する．
- $s(i,a|f) = \frac{1}{m}\sum_{k=1}^{m} \hat{s}_k(i,a|f)$ により，課題表現 i，a の評価値を求める．

7. 評価実験

実験では 2003 年から 2008 年までの間に登録された特許公報のうち，「発明が解決しようとする課題」セクションと，「発明の効果」セクションを持つもの 208,343 件を利用し，提案手法によって課題・特長対を抽出した．

調査対象の技術分野として，実験データに利用した公報の中で最も頻度の高かった，H01L サブクラス（半導体装置，他に属さない電気的固体装置）を利用した（80,533 件）．また，課題表現として，以下の 4 種類を用いた．これらの課題表現は，実験データの課題セクション中で頻度の高いものから人手で選択した．

- 信頼性…低下する (558 件)
- ばらつき…生じる (499 件)
- 歩留まり…低下する (477 件)
- 不良…発生する (354 件)

そして，提案手法を用い，各課題表現 i について $s(i,a|f)$ の高い特長表現上位 30 件を抽出した．このとき，サンプリングを行わずに，分野 f にある全特許明細書を用いて $s(i,a|f)$ を求めた場合と，サンプリングによる繰り返し回数 M を 50, 100, 200 と変化させた場合とを比較し，サンプリングによるノイズ除去の効果を調査した．なお，サンプリングする文書数 n は 5000 に固定した．

ある分野に属する全ての特許明細書から，解決関係にある全ての課題・特長表現対を人手で抽出することは困難であるため，分析対象としている分野の全ての正解を把握することはできない．そのため，評価にあたっては，各課題表現 i との評価値 $s(i,a|f)$ が上位 30 位以内にある課題表現 $\boldsymbol{a}$ について，人手で解決関係にあるか否かを判断し，精度を算出した．

各課題表現ごとの上位 10 件，20 件，30 件の抽出精度を表 1 に示す．課題表現によっては正解数，すなわち解決関係にある特長表現の数が少ないことが考えられるため，精度が低く算出される表現もあることが考えられる．しかし，上位 10 件の抽出精度が上位 20 件・30 件の抽出精度と比較して十分に高くなっていない課題表現があることは，評価値による特長表現の順序付けに工夫の余地があることを示している．

実験で解決関係にある特長表現として適切でないと判定されたものの大半は，以下の二種類のいずれかにある表現であった．

1. 条件や操作を述べている表現など，特長表現として適切でないもの
2. 「低下…起こる」など，解決関係にあるか判定する上で重要な情報が欠けているもの

1 のような誤りの原因としては 6. 節で述べた，記述内容が類似している少数の公報によるノイズが十分に除去しきれていないことが考えられる．従って，ノイズ除去の方法について改良が必要だと言える．また，2 のような誤りは，構文木構造を利用した課題・特長表現の抽出方法に工夫が必要であることを示唆している．

また，サンプリング回数 M ごとの抽出精度を表 2 に示す．課題表現に応じて最適なサンプリングの回数は異なっているが，いずれの表現についても，サンプリングを実施することで

課題表現	Prec@10	Prec@20	Prec@30
信頼性…低下する	0.30	0.35	0.30
ばらつき…生じる	0.20	0.20	0.17
歩留まり…低下する	0.20	0.20	0.17
不良…発生する	0.60	0.5	0.37

表 1: 課題・特長表現対の抽出精度: 課題表現による違い（$M = 200$）

課題表現	M=0	M=50	M=100	M=200
信頼性…低下する	0.17	0.27	0.27	**0.30**
ばらつき…生じる	0.17	**0.33**	0.17	0.17
歩留まり…低下する	0.10	0.10	0.17	**0.17**
不良…発生する	0.27	0.30	0.33	**0.37**

表 2: 課題・特長表現対の抽出精度: サンプリング回数 M による違い（値は上位 30 位の精度 (Precision at 30) を示す．なお，サンプリング無しの結果については $M = 0$ と表記）

精度が向上していることが分かり，ノイズ除去が効果的に働いていることが分かる．

提案手法によって抽出された特長表現の例を表 3 に示す．提案手法により,「信頼性…低下する」と「エロージョン*1…抑制する」のように，構成する単語同士が表層的または意味的に近くないものでも，解決関係スコアが高くなっており，上位 30 位以内に入っていることが分かる．しかし,「歩留まり…低下する」のように構成する単語同士が表層的・意味的に近い特長表現との解決関係しか獲得できていない課題表現もある．本手法によって間接的な解決関係を抽出しやすい課題表現と，抽出しづらい課題表現のそれぞれの特徴については，調査する技術分野や課題表現の数を増やして精査する必要がある．

8. おわりに

本稿では同じ効果を持つ技術の集約を目的として，解決関係にある課題・特長表現の抽出と同定という，新しいタスクを定義した．そして，特許公報の文書構造を利用することにより，解決関係にある課題・特長表現対の抽出方法を提案した．実験の結果，提案手法によって，単語間の関係同定だけでは獲得することのできない，間接的に課題を解決可能な特長表現を抽出可能なことを確認した．

今後の課題として，課題・特長表現の抽出ロジックの改良や，抽出された特許公報から獲得された課題・特長表現対の，技術調査における利用可能性の検証が挙げられる．

課題表現	解決関係にあるとされた特長表現（上位 30 位以内から抜粋）
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

表 3: 獲得された課題・特長表現の例

*1 材料表面の侵食のこと

参考文献

[Bannard 05] Bannard, C. and Callison-Burch, C.: Paraphrasing with Bilingual Parallel Corpora, in *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 597–604 (2005)

[Barzilay 01] Barzilay, R. and McKeown, K.: Extracting paraphrases from a parallel corpus, in *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Vol. 39, pp. 50–57 (2001)

[Dagan 05] Dagan, I., Glickman, O., and Magnini, B.: The PASCAL Recognizing Textual Entailment Challenge, in *Proceedings of the PASCAL Challenges Workshop* (2005)

[Hickl 06] Hickl, A., Williams, J., Bensley, J., Roberts, K., Rink, B., and Shi, Y.: Recognizing textual entailment with LCC 's GROUNDHOG system, in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pp. 80–85 (2006)

[MacCartney 08] MacCartney, B., Galley, M., and Manning, C. D.: A phrase-based alignment model for natural language inference, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pp. 802–811 (2008)

[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 on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access (to appear)* (2010)

[Pantel 06] Pantel, P. and Pennacchiotti, M.: 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)

[坂地 09] 坂地 泰紀, 野中 尋史, 酒井 浩之, 増山 繁：特許文書からのブートストラップ手法を用いた課題・効果表現対の抽出, 情報処理学会研究報告，2009-NL-192 (2009)

[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 10] 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)