大域的・局所的エントロピーに基づいた特許文書中からの効果述語項構造の自動抽出

邊土名 朝飛 [1]　野中 尋史 [1]　河野 誠也 [2]　谷川 英和 [3]

[1] 長岡技術科学大学　[2] 奈良先端科学技術大学院大学　[3]IRD 国際特許事務所
s173348@stn.nagaokaut.ac.jp, nonaka@kjs.nagaokaut.ac.jp,
kawano.seiya.kj0@is.naist.jp, htanigawa@ird-pat.com

1 はじめに

特許情報を分析することは，個人の投資活動，企業における技術開発戦略や M&A，国や地方自治体における政策立案に役立つなど，多大な利益をもたらす．特に特許文書データには重要な新技術の詳細が含まれているため，大きな利用価値がある．しかし，特許文書は技術的な用語が多用されており，また一般的に記述量が多いため，人手で特許文書を読解して分析を実施するためには高度な専門知識と多くの時間が必要となる．そのため，自動的に技術情報を抽出し分析を行うためのテキストマイニング手法が数多く研究されてきた [1].

特許文書には様々な技術情報が含まれているが，その中でも発明がもたらす効果の情報は，その発明が利用者に与える便益，すなわちニーズを示しているため，技術開発戦略や知財戦略の策定において非常に有用である．そのため，発明の効果を抽出する手法 [2, 3] や，抽出した効果を意味的にまとめあげる手法 [4, 5] が研究されてきた．発明の効果に基づいた分析手法のひとつとして，技術-効果パテントマップが挙げられる．技術-効果パテントマップとは，技術シーズを表す発明の技術要素と，その技術要素によってもたらされる効果をそれぞれ軸として特許出願状況を可視化したものであり，ニーズ（効果）とシーズ（技術要素）の両方を加味した特許分析を行うことが可能となる．

本研究では，特許文書中から発明の効果を抽出することを目的として，発明の効果らしい述語項構造（効果述語項構造）を自動的に抽出する手法を提案する．述語項構造とは，述語とその項からなる構造であり，複雑な構造を持つ文章であっても「何を，どうした」といった意味関係を明示的に表現することができる．例えば，「本発明によれば、粘着性物質の付着を防止することができ、メンテナンスを最少限に済すことができる。」という文があったとき，その効果述語項構造は**{メンテナンス [項:ヲ格]，最小限 [項:ニ格]}⇒ 済すことができる [述語]** となる．述語項構造に基づいて効果を抽出することで，単語や句単位で抽出した場合よりも，その効果の意味をより良く捉えることが可能となる．

また，本手法は，述語項構造の発明の効果らしさを，抽出対象となる特許文書集合内での単語エントロピーに基づいて推定する．これにより，**効果 [項:ガ格]⇒ ある [述語]** といった意味のない述語項構造が抽出されるのを防ぎ，かつ分野によって効果の表現が大きく異なっていても対応することが期待できる．

2 関連研究

特許文書から発明の効果を抽出する手法として，「〜ができる。」といった効果を表す文章に頻出する表現（効果手がかり表現）をブートストラップ的に獲得する研究 [2, 3] が挙げられる．また，酒井らの手法で抽出された効果手がかり表現から，直接的に効果を表す語（効果語）をパターンに基づいて抽出する研究 [4] も存在する．しかし，これらのブートストラップに基づいた先行研究の手法は，抽出対象となる文章に記載されている効果の表現が，獲得した手がかり表現と僅かにでも異なっていると効果が抽出できなくなるという問題がある．

一方，谷中らは，特許文書から発明の効果の要点を抽出するために，「本発明は... する」という形式で記述された文章から，述語項構造解析を用いて「本発明」を主語とする動詞と目的語を抽出している [5]．谷中らの方法は，述語項構造に基づいているため，複雑な文であっても発明が技術課題に及ぼす効果を明示的に表現することができ，かつ様々な

パターンの記述に対応できる．しかしながら，効果を示す動詞，目的語が「本発明」を主語としない場合は抽出することができない．また，「ものである」「こととする」といった意味を持たない句を取り除くために，「もの」「こと」「課題」などの形式的内容語 [6] を事前に収集する必要がある．

本研究では，特許明細書中の項目「発明の効果」に含まれる文から最も効果らしい述語項構造を抽出することで，文章の記述形式を限定することなく発明の効果を抽出する手法を提案する．また，本手法では，述語項構造の効果らしさを単語エントロピーに基づいて推定するため，形式的内容語の収集を必要としない．

3 提案手法

特許文書中から発明の効果を抽出するために，特許明細書中の項目「発明の効果」に含まれる文（効果文）から最も効果らしい述語項構造（効果述語項構造）を自動的に抽出する手法を提案する．提案手法の概要を図 1 に示し，手続きの詳細を以下に示す．

Step 1: $L_{t=0} \longleftarrow -\inf$
Step 2:「発明の効果」に含まれている単語から大域的エントロピースコア $H_G(w_i)$ を計算
Step 3: 各効果文に対し述語項構造解析を行い，各効果文から効果述語項構造候補を抽出
Step 4: 効果述語項構造候補に含まれている単語から局所的エントロピースコア $H_L(w_i)$ を計算
Step 5: 効果述語項構造候補集合の効果らしさ L_t を計算
Step 6: $L_t > L_{t-1}$ であればノイズ単語を除去
そうでなければ Step 9 へ
Step 7: $t = t + 1$
Step 8: Step 3 から繰り返す
Step 9: 各効果文から効果述語項構造を抽出

なお，Step 2，4 のエントロピーの計算に関しては 3.1 節で，Step 3 の効果述語項構造候補の抽出に関しては 3.2 節で，ノイズ単語除去に関しては 3.3 節でそれぞれ説明する．

本手法の特徴は，大域的な単語エントロピーで効果語らしさを，局所的な単語エントロピーでノイズ単語らしさを推定し，それに基づいて効果文から効果述語項構造を抽出する点にある．

単語 w_i の大域的エントロピースコア $H_G(w_i)$ は，「発明の効果」に含まれる単語から求める．大域的エントロピースコア $H_G(w_i)$ が高い単語は「発明の効果」に出現しやすい単語であり，すなわち効果語である可能性が高い．したがって，高エントロピーの単語を含む述語項構造も効果を表している可能性が高いと考えられる．しかし，大域的エントロピースコア $H_G(w_i)$ では，「発明の効果」に出現しやすい「もの」「発明」「効果」などの意味を持たない単語のスコアも高くなり，「効果-ある」といった意味を持たない述語項構造を抽出してしまう恐れがある．

このようなノイズ単語を除去するために，局所的な単語のエントロピースコア $H_L(w_i)$ を利用する．単語 w_i の局所的エントロピースコア $H_L(w_i)$ は，各効果文から抽出された効果述語項構造候補に含まれる単語から求める．「効果」といったノイズ単語は，効果述語項構造候補における出現頻度が非常に高いと考えられるので，高エントロピーの単語をノイズ単語として除去していく．

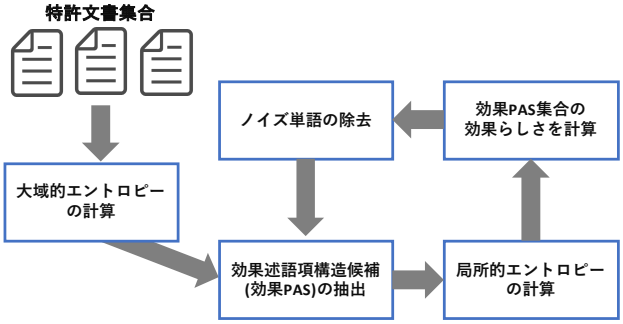

図 1 提案手法の概要図

3.1 大域的・局所的エントロピー

単語 w_i の大域的エントロピースコア $H_G(w_i)$ は，以下の式で求める．

$$H_G(w_i) = -\sum_{d \in D} P_d(w_i) \log_2 P_d(w_i) \quad (1)$$

$$P_d(w_i) = \frac{f_d(w_i)}{\sum_{d' \in D} f_{d'}(w_i)} \quad (2)$$

ここで，D は各特許文書に記述されている「発明の効果」文章の集合，$P_d(w_i)$ は特許文書 d の「発明の効果」内に単語 w_i が出現する確率，$f_d(w_i)$ は特許文書 d の「発明の効果」内に単語 w_i が出現する頻度である．

単語 w_i の局所的エントロピースコア $H_L(w_i)$ は，

以下の式で求める．

$$H_L(w_i) = -\sum_{s \in S_E} P_s(w_i) \log_2 P_s(w_i) \quad (3)$$

$$P_s(w_i) = \frac{f_s(w_i)}{\sum_{s' \in S} f_{s'}(w_i)} \quad (4)$$

ここで，S_E は効果述語項構造候補集合である．効果述語項構造候補は，効果文 1 文から 1 個抽出される (3.2 節)．また，$P_s(w_i)$ は効果述語項構造候補 s 内に単語 w_i が出現する確率，$f_s(w_i)$ は効果述語項構造候補 s 内に単語 w_i が出現する頻度である．

なお，本研究では，名詞，動詞，形容詞，形状詞の単語のみを使用してエントロピーを計算している．

3.2 効果述語項構造候補の抽出

効果文から効果述語項構造候補を抽出するために，述語項構造のスコアリングを行う．スコアリングした後，その効果文内で最も高いスコアを持つ述語項構造を，効果述語項構造候補として抽出する．

ある 1 つの効果文 e に含まれる述語項構造の集合を $S_e = \{s_1, s_2, \cdots, s_i, \cdots, s_N\}$ としたとき，前方から i 番目の述語項構造 s_i の大域的スコア $Score_G(s_i)$ を以下の式で定義する．

$$Score_G(s_i) = \frac{1}{|S_e| - i + 1} \max_{w_j \in \mathcal{V}(s_i)} H_G(w_j) \quad (5)$$

ここで，$|S_e|$ は効果文 e に含まれている述語項構造の個数，$\mathcal{V}(s_i)$ は述語項構造 s_i に含まれている単語の集合である．文中で効果について言及している箇所は一般的に文の後方に位置するため，述語項構造 s_i の後方からの位置の逆数 $\frac{1}{|S_e|-i+1}$ をエントロピースコアにかけることで重み付けしている．

3.3 ノイズ単語の除去

効果述語項構造候補から，**効果 ⇒ ある**といった意味を持たない述語項構造を抽出しないようにするために，ノイズ単語を自動的に除去する．

まず，効果述語項構造候補を抽出（3.2 節）した後，単語の局所的エントロピースコア $H_L(w_j)$ を計算する（3.1 節）．次に，各効果述語項構造候補 s_i の局所的スコア $Score_L(s_i)$ を以下の式で求める．

$$Score_L(s_i) = \max_{w_j \in \mathcal{V}(s_i)} H_L(w_j) \quad (6)$$

次に，効果述語項構造候補集合 S_E の効果らしさを，以下の式で求める．

$$L_t = \sum_{s \in S_E} Score_G(s) - Score_L(s) \quad (7)$$

L_t の値は，効果らしい効果述語項構造候補が多く抽出されている場合には高く，ノイズらしいものが多く抽出されている場合には低くなる．このとき，L_t が 1step 前の L_{t-1} より高くなれば，局所的エントロピースコア $H_L(w_j)$ が最も高い単語をノイズ単語として除去し，再度ノイズ単語除去の処理を行う．この繰り返し処理により，効果語らしい単語だけが残り，効果述語項構造が抽出されやすくなると考えられる．

4 評価実験

4.1 データセット

NTCIR-6 の日本語公開特許公報全文データ[1)]（期間：1993〜2002 年，文書数：3,496,252 件）から，国際特許分類の C22C（合金分野）に属する特許 97 件を選択し，それらに記述されている「発明の効果」の文 234 件を評価用データとして用いた．

正解データとして，上記の 234 件の文に対して効果を記述している箇所にタグ付けしたものを用いた．文内に複数の効果が記述されている場合は，それぞれ独立してタグを付与している．タグ付け作業は 2 名のアノテーターの合議により行われた．

4.2 実験条件

効果が正しく抽出されたかの判定は，2 名の合議に基づいて正否を判断する人手評価を実施した．また，1 文に複数の効果が含まれていた場合，その全ての効果を抽出していれば正解とする評価 (all) と，1 件でも効果が抽出できていれば正解とする評価 (partial) をそれぞれ行った．評価指標は，Precision(P)，Recall(R)，F 値 (F_1) を用いた．各評価指標の式を以下に示す．

$$P = \frac{|C|}{|E|},\ R = \frac{|C|}{|T|},\ F_1 = \frac{2PR}{P+R} \quad (8)$$

ここで，C は正しく抽出できた効果の集合，E は抽出した効果の集合，T は効果が含まれている文の集合である．

本実験では，格フレーム [7] に基づく述語項構造解析ツールの KNP[2)]を使用した．また，単語エントロピーの計算時に用いる形態素解析器には Sudachi[8] を使用し，表記ゆれを吸収するために単語の正規化を行った．

1) http://research.nii.ac.jp/ntcir/permission/ntcir-6/perm-ja-PATENT.html
2) http://nlp.ist.i.kyoto-u.ac.jp/?KNP

比較手法として，効果文の最末尾の述語項構造を抽出する手法 (LAST-PAS) と，坂地らの Cross-Bootstrapping 法 [3] を採用した．

5 実験結果および考察

表 1 に評価結果を示す．提案手法は，all，partial の両方で最も高い F 値を示した．最末尾から述語項構造を抽出する手法を比較すると，all において F 値が 1.5 ポイント，partial において 3.4 ポイント向上している．これは，文末尾に存在する不要な述語項構造が，ノイズ単語が除去されたことにより抽出されなくなったためだと考えられる．

坂地らの手法は，Precision が最も高いものの，効果を抽出するために利用する手がかり表現を含んだ文の件数は 59 文/242 文 (24.4%) に留まっており，Recall は最も低い結果となった．一方，提案手法は述語項構造解析結果に基づいて効果を抽出しているため，様々な記述パターンに対応することができ，結果として Recall と F 値が最も高くなったと考えられる．

表 1 効果抽出の評価結果

	all			partial		
	P	R	F_1	P	R	F_1
坂地ら [3]	**78.0**	28.2	41.4	**84.7**	30.7	45.0
LAST-PAS	35.0	47.6	40.3	45.7	62.2	52.7
提案手法	36.8	**48.5**	**41.8**	49.3	**65.1**	**56.1**

5.1 エラー分析

提案手法 (partial) における誤抽出の分類および件数を表 2 に示す．効果を含む文からの誤抽出では，部分的に失敗のケースが最も多かった．部分的失敗とは，効果を表す一部語句が欠落したタイプの誤抽出であり，例えば「高率放電特性に優れた蓄電池を得ることができる。」という文章から**優れた蓄電池 ⇒ 得ることができる**のみが抽出され，効果として重要な「高率放電特性」が含まれていないケースである．この場合，述語項構造の項に係っている修飾節の取得範囲を広げることで対処することで改善できると考えられる．

一方，効果を含まない文からの誤抽出は，完全に失敗のケースが最も多く，次に非重要効果を抽出したケースが多かった．「発明の効果」において，(1) 最初に発明の構成を示し，次に (2) 重要ではない細かな効果を述べ，最後に (3) 重要な効果を述べる，という記述パターンが存在するが，先述の誤抽出は主に (1)，(2) に属する文から抽出されたものである．この場合，「発明の効果」内の談話構造を考慮することで誤抽出を低減できると考えられる．

表 2 提案手法 (partial) における誤抽出の分類と件数

	効果を含む文	効果を含まない文
非重要効果	1	20
完全に失敗	21	43
部分的失敗	28	-

5.2 ノイズ単語

本手法により自動的に除去されたノイズ単語 29 個を以下に示す．

自動的に除去されたノイズ単語

する, こと, できる, 合金, 製造, なる, 水素, 得る, 吸蔵, ある, 可能, 発明, 金属, 材料, 優れる, 向上, 化合, 有する, 効果, 特性, 用いる, 方法, 提供, コスト, もの, 容易, 形状, 良好, 安価

上記のノイズ単語を見ると，後半に「安価」「コスト」という直接的に効果を表している語が除去されてしまっているものの，その他の単語は「する」「こと」「効果」など，意味を持たない形式的内容語を中心に除去されていることが分かる．また，「合金」「水素」「金属」など，分野特有の単語も除去されている．今回対象としているのは合金分野の特許であり，上記のような分野特有の単語を含む効果，例えば「水素吸蔵合金を製造することができる」が抽出されたとしても分析上有用な情報であるとはいえないため，除去されたのは妥当と考えられる．

6 まとめ

本研究では，特許文書の項目「発明の効果」から，単語の大域的・局所的エントロピーを利用して発明の効果らしい述語項構造を自動的に抽出する手法を提案した．合金分野の特許を対象に評価実験を行った結果，提案手法は 2 種類の比較手法よりも高い F 値を示した．今後の課題として，抽出した効果述語項構造をクラスタリングし，効果をまとめ上げることが挙げられる．

参考文献

[1] Longhui Zhang, Lei Li, and Tao Li. Patent mining: A survey. *SIGKDD Explor. Newsl.*, Vol. 16, No. 2, p. 1–19, May 2015.

[2] 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.

[3] 坂地泰紀, 野中尋史, 酒井浩之, 増山繁. Cross-bootstrapping : 特許文書からの課題・効果表現対の自動抽出手法. 電子情報通信学会論文誌. D, Vol. 93, No. 6, pp. 742–755, 2010.

[4] H. Nonaka, A. Kobayahi, H. Sakaji, Y. Suzuki, H. Sakai, and S. Masuyama. Extraction of the effect and the technology terms from a patent document. In *The 40th International Conference on Computers Indutrial Engineering*, pp. 1–6, 2010.

[5] 谷中瞳, 大澤幸生. 特許文献を利用した技術課題の抽象化方法の検討. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 1J32–1J32, 2016.

[6] 滝川和樹, 山本和英. 構文片の改良と評判分析への適用. 言語処理学会第 17 回年次大会発表論文集, pp. 111–114, 2011.

[7] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, Vol. 14, No. 4, pp. 67–81, 2007.

[8] Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida, and Yuji Matsumoto. Sudachi: a japanese tokenizer for business. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, 2018.