製品特徴に基づく製品発表プレスリリースの関連特許自動判定

中山優輝　酒井浩之　永並健吾

成蹊大学 理工学部 情報科学科
us182093@cc.seikei.ac.jp {h-sakai, kengo-enami}@st.seikei.ac.jp

概要

本研究では，製品発表プレスリリースと関連のある特許を自動的に判定する手法を提案する．特許明細書から BERT[2]を用いて「発明を使用することによりもたらされる効果が記述してある表現」(以下，重要文）を抽出し，その後，特定の文末表現を用いることで重要文から「製品の特徴を示すキーワード」（以下，重要語）を抽出する．これにより製品特徴の表現とは無関係な文と語の除去を図り，プレスリリースの製品と関連のある特許の検索精度向上を目指す．

１．はじめに

特許は企業の原動力となる新技術やビジネスの大事な指標の一つであり，他社における先行特許を調査することは極めて重要である．またそのような調査では，単に開発製品と関連のある技術分野の特許を調査対象とするだけでなく，競合している他社の特許と製品の双方に着目し，より細部での関連性を対象として特許を特定することが必要とされている．しかしながら，文書から製品の機能に関する特徴的な技術を十全に把握し，かつ膨大にある特許明細書から当該製品と関係のある特許を見つけ出すことは容易ではない．

そこで本研究では，特許明細書と製品発表プレスリリースの双方から，重要文と重要語を抽出することで，対象としたプレスリリース記事と関連のある特許明細書を自動的に判定することを目的とする．例えば，「カメラ」の製品発表プレスリリースを入力とした場合，「レンズ」や「画像素子」に関する特許が，関連性があると自動的に判定され，関連特許検索に応用可能な技術である．

本研究により，発表された製品に導入された技術に関する特許を素早く検索することが可能となり，競合している他社製品の特許を簡単に把握することができるようになる．

関連研究に，関連特許検索を二値分類問題として扱った文献[1]があるが，こちらは BERT[2]が用いられているものの，対象にはノイズとなる語も含まれており，正例負例が不均一となる特許検索を主とした言語モデルには適していない．また，特許文献における発明の作用・効果を抽出する手法[3]があり，こちらは対象とは無関係とされる文や句の除去が人手で行われ，最終的に残った文章について SVM を使用することで重要文抽出を試みていた．同様に特許の機能表現に着目した類似特許検索法が提案された文献[4]では，相互情報量を使用することで文章を短文に分割し，語の従属関係に基づいた類似度計算が行われている．先の両文献についてはいずれも BERT を使用した重要文の自動抽出が試みられていない点で本研究とは異なっている．

その他，関連研究として文献[5]があるが，当該文献では関連度計算に用いられる特許文の対象が人手によって絞り込まれており，特許のもたらす効果とは無関係な記述が多く存在していた．また，関連度計算においては文書の分散表現が用いられておらず，語の意味関係が考慮できていない．それに対して本研究では，特許明細書と製品発表プレスリリースの双方から BERT を使用して重要文を抽出することでノイズとなる文章を除去し，Word2Vec を使用した文書の分散表現を用いることで語の意味関係を考慮した関連特許検索手法を提案する．

２．提案手法

本提案手法は以下の 3 つの Step で構成される．

Step1: BERT を用いて，特許明細書及びプレスリリース記事からそれぞれ重要文を抽出

Step2: Step1 で抽出された重要文から重要語をそれぞれ抽出

Step3: Step2 で抽出された重要語を用いて，プレスリリース記事の関連特許を検索

本研究では 2017 年度に受理された特許データ 446,644 件，及び日経プレスリリース記事の「情報・通信」と「素材・エネルギー」に関する 16,180 記事を使用する．

2.1. BERT を用いた重要文抽出

BERT を用いることで，例えば「本発明によれば、ステイプル処理やパンチ処理等の後処理を伴う印刷において印刷において試し印刷をした場合における記録紙の無駄を削減することができる」などの文を抽出し，「無駄を削減する」といったような『発明を使用することによりもたらされる効果』の記述を重要文として抽出する．また，重要文を抽出することで製品特徴とは無関係なノイズとなる文の除去が可能となる．

2.1.1. BERT の学習データ

BERT の学習データは全 8 種ある特許 IPC について 1 種 1,500 件として無作為に選んだ計 12,000 件の特許データから，それぞれ正例・負例 10,000 文を無作為に抽出し，計 20,000 文を使用する．

正例に関しては「発明を使用することによりもたらされる効果が記述してある表現」が特許における「発明の効果」（advantageous effects）に記載されている文と同義であるという仮定に基づき無作為に抽出を行った．また，負例に関しては「重要文とは無関係な文」が特許における「実施例」（embodiments example）に記載されている文と同義であるという仮定に基づき無作為に抽出を行った．

2.1.2. BERT を用いた特許の重要文抽出

特許明細書の重要文抽出において使用する BERT モデルは 2.1.1.で作成したデータを学習データとしたモデルとし，テストデータを企業ごとの特許明細書とする．表 1 にテストデータとして用意した企業とその特許データ数の一部を示す．

表 1　各企業と特許データ数（テストデータ）

企業名	特許データ数
富士ゼロックス株式会社	1,727 件
富士電機株式会社	1,316 件
日本電信電話株式会社	1,912 件
ソニー株式会社	2,026 件
株式会社村田製作所	1,768 件
大日本印刷株式会社	2,021 件

2.1.3. BERT を用いたプレスリリース記事の重要文抽出

プレスリリース記事には「素材」や「厚み」といった商品の仕様が箇条書きとなっている文章や，製品とは無関係な「開発企業の説明」などのノイズとなる文章が含まれている．そのため，特許と同様に BERT を用いることで重要文を抽出し，ノイズ文の除去を行なう．プレスリリース記事の重要文抽出において使用する BERT モデルは 2.1.2.で使用したモデルと同一とし，テストデータを表 1 で示された企業に関するプレスリリース各 1 記事とする．

2.2. 文末表現に着目した重要語抽出

次に，BERT によって抽出された重要文から重要語を取得する．例えば，「お客様の投資負担軽減及び売電事業に貢献します。」という文から「貢献します。」という文末表現に係っている文節に含まれる『売電事業』や『投資負担軽減』を取得することで重要語を抽出し，さらに『お客様』といった製品の特徴とは無関係な語の除去を行なう．こうして，重要文からさらに重要語を抽出することで精度の向上を図る．

文末表現の取得について，特許明細書からは BERT の学習データである 12,000 件の「発明の効果」に関する文章 58,959 文を使用し，プレスリリース記事からは「情報・通信」と「素材・エネルギー」に関する 16,180 記事について BERT を用いて抽出された 99,201 文を使用した．また，プレスリリース記事の文抽出に用いた BERT モデルは 2.1.2.で使用したモデルと同一とした．重要語の取得は以下の 3 つの Step で構成される．

Step1: 人手で選定された文末表現から，文末表現に係っている文節に含まれる語を抽出

Step2: Step1 で抽出された語が含まれる文節が係っている文末表現を再取得

Step3: Step2 で取得された文末表現を用いて，特許明細書，および，プレスリリース記事から重要語を抽出

以下，Step1 の具体例を図 1，Step2 の具体例を図 2 に示す．図の赤字は重要語，青字は文末表現をそれぞれ表す．

このような酸素拡散層は酸素コレクターに捕捉された酸素を広い範囲に拡散し、タイヤ外部への酸素の排出を促進する。

※文末表現「促進する。」を選定した場合，
「酸素拡散層」「拡散」「排出」が抽出される．

図 1　選定した文末表現からの語の抽出例（Step1）

酸素コレクターがカーカス層の巻き上げ部に対応する部位に埋設される場合、酸素コレクターよりもタイヤ外表面側で酸素輸送コードと近接する位置に織布又はメッシュを含む酸素拡散層を設けることが好ましい。

※Step1 で抽出された「酸素拡散層」から，
新たに文末表現「好ましい。」を取得する．

図 2　Step1 で抽出された語から文末表現を再取得（Step2）

2.3. プレスリリース記事の関連特許検索

2.2.で抽出された重要語に Word2Vec を使用することで各文書（特許明細書，および，プレスリリース記事）を固定長のベクトル（分散表現）として表現する．これにより文書間で同一名詞が出現していない場合でも類似度計算を行い，関連文書を検索することができる．文書の分散表現は，各重要語の分散表現を用いて式(1)で求める．

$$DocVec = \frac{1}{n}\sum_{i=1}^{n} WordVec_i \qquad (1)$$

$WordVec_i$: 文書に含まれる重要語 i の分散表現
n: 文書内に含まれる重要語の数

プレスリリース記事 D と特許明細書 T 間の類似度計算にはベクトル空間モデルに基づいて，コサイン類似度を使用する．

3．評価

重要文抽出に使用する BERT については，バッチサイズを 128，エポック数を 5 として実験を行なった．また，文末表現・重要語抽出において形態素解析器は MeCab を使用した．ここで文末表現については閾値を DF 値 5 以上，IDF 値 2 以上のものを抽出し，重要語についての DF 値，IDF 値の閾値は文末表現同様，かつ「英数字」「記号」が含まれている語を除いた 2 文字以上の複合名詞として抽出を行なった．また，文書の分散表現を求めるうえで使用する Word2Vec は約 2GB の Wikipedia データを用いて作成されたモデルとした．なお，作成した Word2Vec のモデルは，複合名詞を 1 つの語として扱ったモデルであり，複合語の分散表現を得ることができるものである．

上記によって，文末表現については最終的に特許から計 4,182 個，プレスリリース記事から計 1,816 個が取得された．

評価における比較手法として，特許明細書・プレスリリース記事 8 記事（P1～P8）に対して重要文抽出を行わずに類似特許を求めた場合と，重要文抽出を行った場合，提案手法である重要文抽出と文末表現を使用した検索結果それぞれ上位 10 件及び上位 20 件における精度を表 2 に示す．

表 2　各プレスリリース記事の類似特許検索精度

プレスリリース記事	重要文抽出なし	重要文抽出あり	重要文抽出＋文末表現
P1	20% / 25%	30% / 35%	50% / 50%
P2	10% / 15%	40% / 25%	80% / 40%
P3	60% / 50%	60% / 50%	60% / 60%
P4	70% / 50%	60% / 40%	80% / 55%
P5	20% / 30%	30% / 40%	30% / 45%
P6	10% / 10%	70% / 60%	90% / 85%
P7	40% / 20%	40% / 20%	40% / 20%
P8	70% / 60%	70% / 65%	90% / 70%

（上位 10 件の精度 / 上位 20 件の精度）

P1: 富士ゼロックス、欧州医療機器規則に対応する取扱説明書の多言語翻訳サービスを提供開始
P2: 富士ゼロックス、複合機アプリケーション「クラウド連携アプリケーション for DX Suite」を提供開始
P3: 富士電機、太陽光発電設備のコスト削減を実現する蓄電池併設型マルチ PCS を発売
P4: 富士電機、第 7 世代「X シリーズ」IGBT モジュールの系列を拡大－1,700V 耐圧製品のサンプル出荷を開始
P5: NTT、秘密計算システム「算師」の試用提供を開始－大切なデータを安心・安全に利活用
P6: ソニー、大口径超広角ズームレンズ G マスター「FE12-24mmF2.8GM」を発売

P7: 村田製作所、サイバートラスト社のセキュアIoT プラットフォーム対応の MCU 内蔵 Wi-Fi モジュールを商品化
P8: 大日本印刷、評価分析機能付きデジタルテストシステムを開発

また，例として，P6 における類似特許の検索結果上位 20 件を図 3 に示す．ここで，図の赤字は，評価の結果，プレスリリース P6 の関連特許として適していると判定したものである．

赤色は類似特許

1 レンズユニット、撮像装置、および制御方法	13 情報処理装置、位置および／または姿勢の推定方法、およびコンピュータプログラム
2 レンズユニット、撮像装置、および制御方法	14 画像処理装置、画像処理方法及びコンピュータプログラム
3 表示装置及び表示制御方法	15 電子機器
4 制御装置、制御方法および露光制御システム	16 表示装置及び表示制御方法
5 複眼撮像装置	17 情報処理装置及び情報処理方法、並びに画像表示システム
6 フィルタ制御装置およびフィルタ制御方法、ならびに撮像装置	18 撮像装置、固体撮像素子、カメラモジュール、電子機器、および撮像方法
7 撮像制御装置、撮像装置、及び撮像制御方法	19 医療用立体顕微鏡光学系及び医療用観察装置
8 撮像装置、および電子装置	20 投射型表示装置
9 広角レンズおよび撮像装置	
10 信号処理方法、及び撮像装置	
11 制御装置および制御方法	
12 撮像装置および撮像レンズ	

図 3　P6 における類似特許の検索結果

4．考察

全体のプレスリリース記事に関して，提案手法である「BERT における重要文抽出」と「文末表現を用いた重要語抽出」を行なうことで精度は最大 90%，精度向上率については重要文抽出を行なっていないものと比較して最大 80%の向上を示した．

また，プレスリリース記事 P2・P7 については，上位 20 件の精度が上位 10 件の精度の半分となっているが，これは類似特許が上位 10 件のうちに偏っているためである．前提として，本研究では 2017 年度に受理された特許を企業ごとに対象としているため，1 つのプレスリリース記事に類似する特許が一定数以上あるとは限らない．そのため，精度は低くとも類似特許検索としては十分な性能であると考えられる．

また，P5 については他の記事と比較して提案手法における精度の向上があまり測れていない．ここで，図 4 にて P5，図 5 にて P6 で抽出された重要語を記す．

利活用,日本電信電話,日本電信電話株式会社,昨今,イノベーション,変化,トランスフォーメーション,デジタル,要因,解消,運用,利点,開示,データ運用,改良

図 4　P5 で抽出された重要語

軽量,小型軽量,マスター,小型,ならでは,描写,撮影,堅牢性,堅牢,迫力,画角,被写体,最先端技術,光学,最先端,設計,光学設計,調整,像面,レンズ,徹底的,湾曲,コーティング,フレア,ナノ,大幅,ズーム,距離,ズーム全域,近接,搭載,能力,全域,リニア,モーター,リニアモーター,性能,インナーフォーカス,インナー,フォーカス,変動,好み,合わせ,マニュアル,リング,フォーカスリング,ピント,配慮

図 5　P6 で抽出された重要語

プレスリリース記事 2 記事で抽出されたそれぞれの重要語を比較すると，抽出された重要語の多い P6 のほうが，重要語があまり抽出されていない P5 よりも表 2 において精度が劣っていることが分かる．P5 で抽出された重要語が少ない原因として，2.2.にて獲得された文末表現が P5 の記事文中に十分に含まれていないことが挙げられる．そのため，重要語を抽出できる文末表現の取得数を増やす手法が必要であると考える．

5．むすび

本研究では，製品発表プレスリリースと関連のある特許を自動的に判定する手法を提案した．評価の結果，特許明細書・プレスリリース記事の双方について「BERT を用いた重要文抽出」並びに「文末表現を用いた重要語抽出」を行なうことで，類似特許検索の精度向上が見込めることを示した．

しかしながら，一部のプレスリリース記事において重要語の抽出が十分でなかったことから，満足される精度を示さない記事も見受けられることが分かった．今後は適切な文末表現の取得数を増やすことで，類似特許検索において有用に活用できるような手法を検討していくことが求められる．

参考文献

[1] Dylan Myungchul Kang, Charles Cheolgi Lee, Suan Lee, Wookey Lee, “Patent Prior Art Search using Deep Learning Language Model” , IDEAS '20: Proceedings of the 24th Symposium on International Database Engineering & Applications（2020）.

[2] Devlin，Jacob，et al，“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.”，arXiv preprint arXiv:1810.04805 (2018).

[3] 原田綾花，太田貴久，小林暁雄，増山繁，野中尋史，酒井浩之，“特許文書からの発明に関する特徴的技術とその効果の抽出”，言語処理学会第１９回年次大会発表論文集，pp512-515，2013.

[4] Jian-Hong Ma, Ning-Ning Wang, Shuang Yao, Zi-Mo Wei, Shuai Jin, “Similar Patent Search Method Based on a Functional Information Fusion”, IDEAS '20: Proceedings of the 24th Symposium on International Database Engineering & Applications（2018）.

[5] 酒井浩之，増山繁，“製品特徴に基づく製品発表プレスリリースと特許との関連性の判定”，言語処理学会第１９回年次大会発表論文集，pp725-728，2013.