メタデータと単語分散表現素性に着目した特許文書の発明者推定

安齊和音　　青野雅樹
豊橋技術科学大学　情報・知能工学課程
anzai@kde.cs.tut.ac.jp, aono@cs.tut.ac.jp

1. はじめに

企業における知財戦略において特許取得は，自社の製品を保護するうえで非常に重要である．特許出願にかかるコストは，1 件につき 30~50 万円と，決して容易に出願できるものではない．そのため，企業における知的財産部では，特許出願を拒否されないために類似特許を精査する必要がある．しかし，特許出願の数は年間約 30 万件のペースで増えていくため，総当たりで類似特許を検索することは困難になってきている．一般的に特許には同じ発明者は似たような語彙や言い回しを用いた特許を多く出願する傾向がある．類似特許であっても，「遊技機」「ゲーム機」のように筆者により言葉使いが異なることがある．

本研究では，その傾向に着目し，特許の言葉使いの違いが吸収できるように統語的な素性に代表されるメタデータと単語分散表現素性に着目し，特許発明者を推定する手法を提案する．

2. 関連研究

特許に関する研究はさまざまな分野で行われている．鈴木ら[1]，小西ら[2]は，特許文書における重要単語の抽出に関する研究を，福田ら[3]，安藤ら[4]は，特許文書から最新の技術動向を読み取る研究を行っている．

一方でニュース記事やレビュー記事などの著者推定タスクも近年，活発に研究が行われている． 著者推定タスクでの精度を競う PAN@CLEF[5]では，Houda ら[6]，Bagnall[7]等によってさまざま手法を用いて著者推定の精度向上を目指している．

特許をその中に現れるトピックで分類する試みとしては，Venugoplan らの研究[7]が知られている．特許から，今後出そうな新しい製品を予測する研究としては，Lee ら[8]や Livotov[9]の研究が報告されている．特許の傾向から，今後の特許を予測する研究は，「パテントマップ」として，古くから研究されている（例 野崎 [10]）．NCTIR で始まった各種の特許情報処理(検索，分類，分析，翻訳)は藤井ら[11]にまとめられている．

特許文献における発明者推定の研究は，我々が知る限り，まだ存在しない．そこで，本研究では，発明者推定に焦点をあて，その精度向上を目指すべく研究に取り組んだ．

3. データセット

本研究では，独自に収集した特許データ 3180 件を用いて実験を行う．このデータセットには，一人当たり 20 件，合計 159 人の特許全文が含まれており，実験ではこれを発明者あたり 3:1 にランダムに分割し，それぞれ訓練データ・テストデータとして扱う．このデータセットを特許文書の分類基準となる IPC(International Patent Classification)の頭 1 文字で分類した内訳を図 1 に示す． また，特許に付属する IPC の例を図 2 に示す．

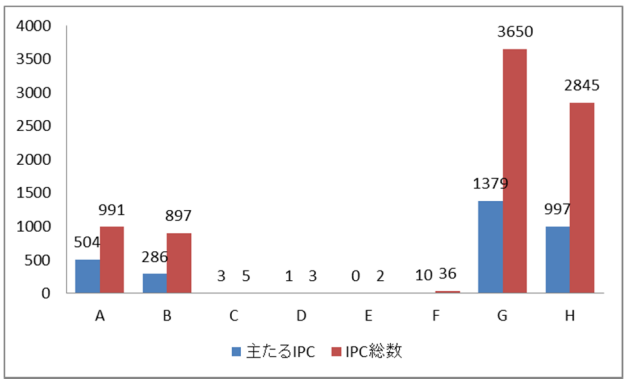

図 1 データセット内訳(IPC 別)

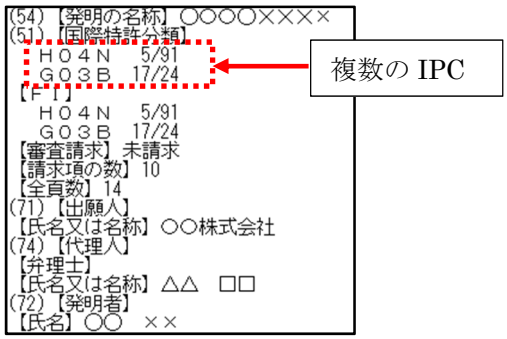

図 2 特許 IPC 付与の例(国際特許分類)

図 1 において青色の棒グラフはその特許文書が保有する IPC の内，主となる IPC の数を示す．日本における特許文書において IPC は国際特許分類と表記される．図 2 を見るとわかる通り，複数の IPC が特許文書に付与されていることがわかる．主たる IPC とは図 2 において国際特許分類の先頭にくるものを表す．また，図 1 において赤色の棒グラフは特許文書全てに付属する IPC の数を示す．

本研究で用いたデータセットは主に，情報処理装置，画像形成装置，プログラム，ゲーム機，制御装置，

装置の操作方法, 等の特許文書から構成される. そのため, 図 1 に示したように, 情報処理装置, プログラム, 画像形成装置が IPC に含まれる G が最も多く, 制御装置が含まれる H, 装置の操作方法に関する特許が含まれる B, ゲーム機に関する特許が含まれる A が次点で多くデータセットに入っている.

上記に記述した特許データセットの他に, 3.2 節で記述する単語分散表現のモデルのために日本語 Wikipedia[12]と, NTCIR [13]の約 300 万件の特許データのうち, 著者推定用のデータとは独立に IPC を有する約 30 万件の特許データを使用する.

4. 提案手法

以下では我々が着目した, 特許のメタデータと単語分散表現素性による特許発明者の推定手法に関して, 4.1 節で提案手法の概要を述べた後, 4.2 節では, データセットの前処理について, 4.3~5 節で各提案手法について説明する.

4.1 提案手法の概要

本研究では, 3 種類の手法により発明者推定を行う. 第一の手法は, Bag of Words 素性と, 観察に基づいて考案したメタデータ素性を SVM (Support Vector Machine)を用いて分類する手法である. 第二の手法は, Word2Vec 学習済みモデルを初期重みとした Embedding と, 1 次元畳込みによる文章における単語の周辺の語を考慮した特徴を用いた推定手法である. 最後に, 上記 2 つの手法を掛け合わせ, 両方の特性を活かし, 精度の向上を狙った Ensemble な推定手法である.

4.2 データセットの前処理

データセットに含まれる特許の構成の一例を図 3 に示す.

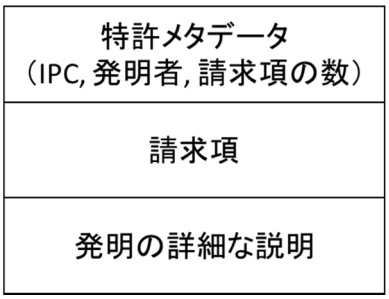

図 3 特許文書の構成例

特許文書は一般的に, 発明者の発明の情報を元に弁理士が代理で執筆する. そのため特許文書は図 3 のように, 特許のメタデータが記載された「特許メタデータ」部分, 「請求項」部分, 「発明の詳細な説明」部分の大きくわけて 3 つの特徴から構成される.本研究では「特許メタデータ」部分と「発明の詳細な説明」部分を使用する.

4.3 Bag of Words 素性+提案素性

この節では, 特許メタデータ素性を用いた機械学習による特許文書の発明者推定手法を提案する. 後述する 4.3.1 から 4.3.3 で得られる素性ベクトルを連結したものを提案素性ベクトルとする. 各素性ベクトルの次元数は表 1 の通りである. 学習モデルは SVM を用いた.

また, Gini 係数(ランダムフォレスト)による重要度測定を SVM による推定と独立に行い, それぞれの素性の重要度を確認した.

4.3.1 Bag-of-Words

特許文書内「発明の詳細な説明」から名詞のみを抽出し, その頻度をデータセットにおける最大の単語頻度で割ることで 0~1 の実数に正規化したものを素性とする. 文書における単語の分かち書きと品詞解析は形態素解析ツール MeCab[1] を用いる.

4.3.2 統語的な素性

統語的な素性として, 動詞の数, 形容詞の数, 副詞の数, 連体詞の数, 名詞の数, アルファベットの数, 文章の長さ, タイトルの長さ を用いる. それぞれの素性はその素性の最大の値で割ることで正規化する.

4.3.3 メタデータ素性

特許文書における特有の素性として, 以下のメタデータを素性として提案する. メタデータは, 請求項の数, 図の数, 特許文書それぞれに割り振られている IPC の頭 1 文字の数の 3 種類を提案する. なお, それぞれの素性はその素性の最大の値で割ることで正規化する.

表 1 各素性ベクトルの次元数

素性名	次元数
Bag of Words（TFモデル）	9039
特許文書の長さ	1
タイトルの長さ	1
アルファベットの出現数	1
各品詞の数 (動詞,形容詞,副詞,連体詞,名詞)	5
図の数	1
請求項の数	1
各IPCの数	8

[1] http://taku910.github.io/mecab/

4.4 Embedding+1 次元畳込み

本節では，深層学習における Embedding と 1 次元畳込みを用いた特許文書の発明者推定手法を提案する．図 4 に本研究で用いる深層学習モデルのネットワーク構造を示す．

まず，訓練データの特許文書に現れる語彙の頻度上位 2 万単語を抽出し，インデックス化を行う．そのインデックスを用いて訓練データの文章に現れる単語を 200 次元の重みベクトルに変換する Embedding 処理を行う．Embedding 処理の際，初期重みを定める必要があるが，本研究ではあらかじめ学習した 200 次元の Word2Vec モデルを用いる．

続いて，Embedding 処理によって変換された単語で構成された訓練データの文章の周辺語を考慮しつつ畳込みを行う．本手法ではフィルター数を 256，カーネルサイズを 5 とした．具体的には，まず訓練データの特許文書中の単語を，現れる順番にターゲット（中心単語）とする．ターゲットとなる単語の前後 5 語(ターゲット含む)から特徴マップを抽出し，256 種類のフィルターをかけて表現する．図 5 にカーネルサイズ 5 の 1 次元畳込みの処理を表した様子を示す．

その後，複数ある特徴マップに対して最大プーリングを行い，訓練データの特許文書 1 件あたり 1 つの特徴を定め，全結合層において softmax を用いてマルチクラス分類問題として著者推定を行う．

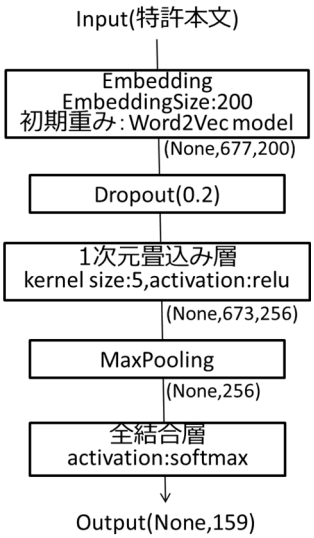

図 4　ネットワーク構造

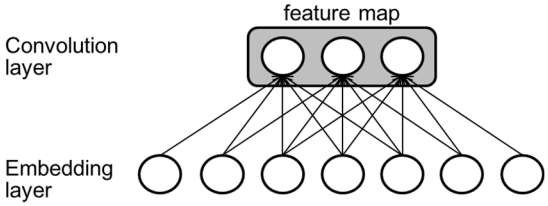

図 5　1 次元畳込み処理

本手法で Embedding 処理に用いる初期重みの Word2Vec モデルは，日本語 Wikipedia と NTCIR 特許データのそれぞれ 200 次元で表されたモデル 2 種類を使用し，比較する．NTCIR 特許データは，NTCIR 特許データセット 1993~2002 年のうち，2002 年に公開された特許約 30 万件の特許全文を，形態素解析ツール MeCab を用いて分かち書きしたものである．本手法ではバッチサイズを 64,エポック数を 80 として実験を行った．深層学習のフレームワークとして Keras を用いた．また，本手法では各データの文章の長さをおおよそ均一にするため，特許文書内「発明の詳細な説明」の先頭 2000 文字を使用した．

4.5 Ensemble Learning

4.3 節, 4.4 節から得られるクラス確率の要素積をとるアンサンブル学習を行う手法を提案する．最終的な分類結果はそれぞれのクラス確率の要素積をとった結果から最も高い発明者のクラス確率を選ぶことで，その特許文書の発明者とする．図 6 にアンサンブルされるまでの流れを記載する．

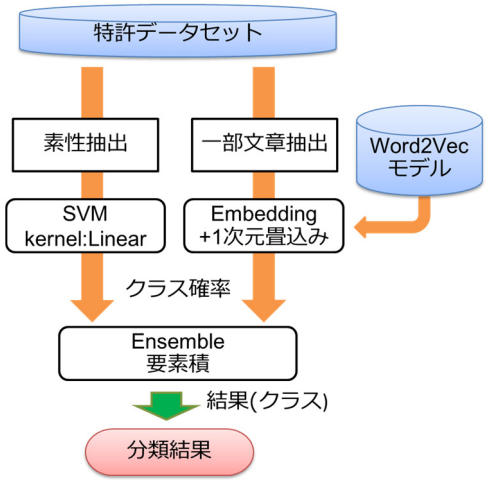

図 6　Ensemble 学習の流れ

5. 評価実験

ベースラインとして 4.3 節で用いた SVM の素性を Bag of Words のみで実験したものを用いる．分類精度を評価する指標として accuracy を用いる．また 4.4 節で述べた深層学習モデルの損失関数に categorical_crossentropy を用いる．

5.1 素性の重要度

4.3 節において提案した素性の重要度の確認を行う．重要度の確認にはランダムフォレストを用いる．Bag of Words によって得られた単語素性を含めた重要度グラフを図 7 に，提案素性のみの重要度グラフを図 8 に示す．

図 7 において下線が引いてあるものが提案素性で

ある. IPC が A である数,という素性が重要であることがわかる. 一方, 他の IPC では重要度が低く, 効果がなかった. これは本データセットでは工業製品や制御機器, プログラム等のいわゆる工学系の特許が多いが, その中には「ゲーム機」や「遊技機」のアミューズメントに関する特許も含まれており, これらは IPC が A に分類されるため, 「IPC が A の数」という素性の重要度が高くなったと考えられる.

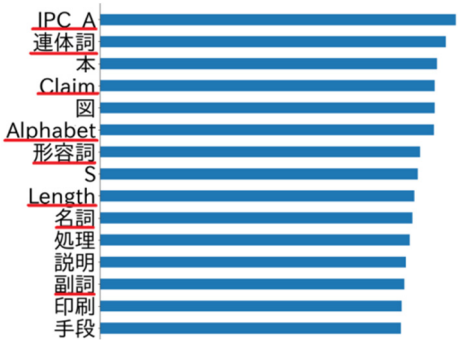

図 7　素性の重要度(BoW+メタデータ素性)

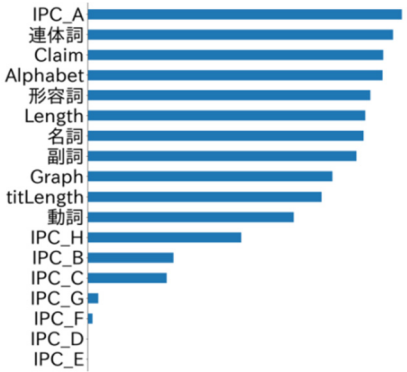

図 8　素性の重要度(メタデータ素性のみ)

5.3　実験結果

ベースラインと各手法と比較した表を表 2 に示す.

表 2　各手法の分類精度

method	accuracy
ベースライン	0.5698
提案手法①	0.5836
提案手法②(日本語Wikipedia)	0.5145
提案手法②(NTCIR2002特許データ)	0.6855
提案手法③	0.7509

① BoW (重み付け:TF)+メタデータ素性
② Embedding→ 1次元畳込み
③ Ensemble Learning

Ensemble した結果が, それぞれの手法に対して精度が向上していることが分かる. これは BoW とメタデータ素性による効果と 1 次元畳込みによる効果が合わさったことが要因であると考えられる. また, Word2Vec モデルに特許データ学習したものを使う場合のほうが, 日本語 Wikipedia を学習したものを使うより精度が向上していることがわかる. これは特許に用いられる専門用語が日本語 Wikipedia には含まれず NTCIR 特許データを学習したものでは含まれているためだと推測できる.

6.　おわりに

本研究では, メタデータと単語分散表現素性に着目した特許文書の発明者推定手法を提案した. 評価実験ではベースラインに比べ精度が向上した.

今後の課題は, 更なる精度向上や新しい素性の提案などが挙げられる.

謝辞

本研究の一部は, 科研費基盤（B）（課題番号 17H01746）の支援を受けて遂行した.

参考文献

[1] 鈴木祥子ら, 特許請求項からの新規性に関わるキーワード抽出 (言語理解とコミュニケーション) -- (第 9 回テキストマイニング・シンポジウム). 電子情報通信学会技術研究報告 2016

[2] 小西慶和ら, 特許文書からの技術動向調査に有効な技術用語の抽出. 第 14 回情報学ワークショップ 2016

[3] 福田 悟志ら, 論文と特許からの技術動向情報の抽出と可視化 情報処理学会論文誌データベース（TOD） 6(2), 16-29, 2013

[4] 安藤 俊幸ら, 機械学習を用いた効率的な特許調査, 情報プロフェッショナルシンポジウム予稿集 2017(0), 83-88, 2017

[5] Houda Alberts, Author Clustering with the Aid of a Simple Distance Measure—Notebook for PAN at CLEF 2017

[6] Douglas Bagnall, Authorship Clustering Using Multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2016

[7] Subhashini Venugopalan and Varun Rai, Topi based classification and pattern identification in patents, Technological Forecasting & Social Change, Vol. 94, pp.236-250, Elsevier, 2015

[8] Changyong Lee, Ohjin Kwon, Myeongjung Kim, and Daeil Kwon, Early identification of emerging technologies: A machine learning approach to using multiple patent indicators, Technological Forecasting & Social Change, Vol. 128, pp. 291-303, 2018

[9] Pavel Livotov, Using patent information for identification of new product features with high market potential, Procedia Engineering, Vol. 131, pp. 1157-1164, 2015

[10] 野崎篤志「特許情報分析とパテントマップ作成入門」発明推進協会, ISBN 978-4827111644, 2011

[11] 藤井　敦, 谷川英和, 岩山真, 難波英嗣, 山本幹雄, 内山将夫,「特許情報処理 : 言語処理的アプローチ」コロナ社、ISBN 978-4339-027556, 2012

[12] 日本語 Wikipedia エンティティベクトル http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector

[13] NTCIR patent datasets 1993-2002 http://research.nii.ac.jp/ntcir/data/data-ja.html