テキスト情報とメタ情報の双方を活用した特許文書分類

Patent document classification using both text and meta information

株式会社 NTT データ数理システム　データマイニング部グループリーダー・主任研究員
岩本　圭介

1999 年株式会社数理システム(現:株式会社 NTT データ数理システム)入社。自然言語処理や深層学習に関わるツール・手法開発及び分析業務に従事。現職はデータマイニング部グループリーダー・主任研究員。

iwamoto@msi.co.jp

1 はじめに

文書を自動的に定められたカテゴリへと分類することは、自然言語処理及び機械学習技術双方の応用分野として大きな関心が寄せられており、その有用性も極めて高いものになっている。例としては

- 蓄積された不具合情報の内容から、問題が発生した部品・部位の種別や発生事象をカテゴライズし、それらの間の関連を把握する
- Web から採取した個人の声をポジティブ·ネガティブに分類して、製品やサービスが世の中にどう受け止められているかを知る

など広い分野における要請があり、特許文書が対象である場合でもこれは例外ではない。過去の Japio YEAR BOOK においても、機械学習を利用した特許のカテゴリ分類については[1], [2]で論じられており、また特許検索の効率化のため「正解文書」群を選り分けることについては[3]で言及されている。

さらに、文書には、その内容の中心であるいわゆる「文章(テキスト)」の部分の他に、様々な属性情報(メタ情報)が付随していることが多くある。こういった情報もカテゴリ分類を行う際の手がかりとすることができれば、テキストだけでは分類性能が思わしくない、メタ情報だけでは判断がつかない、というようなケースにおいてもそれらを相互補完的に利用して精度を向上させられる可能性がある。先に挙げた２例と紐づけて考えると

- 不具合情報の報告内容テキストに加えて、発生場面や対応人員、機器設置状況といった属性情報もあわせて事象のカテゴライズに利用する
- 同じようなことを述べているがその解釈は個人の背景事情に依存することもあるため、個人のプロフィール情報も利用して Web 上の意見を分類する

といった拡張が考えられる。

本稿では、テキスト情報とメタ情報とを同時に扱って文書の分類を行う手法を紹介し、特許文書のカテゴライズを試みる。以降、２章で手法の解説を行い、３章で実験の結果を述べる。最後に、４章で実験に用いたツールの解説とまとめを行う。

2 テキスト情報とメタ情報

2.1　テキスト情報とメタ情報のベクトル化

特許文書において、課題・解決手段・請求項といった文章で記述情報と、公開日・分類コード・出願人や発明者といった書誌情報との間の関係は、そのままテキスト情報とメタ情報との関係にあたる。一般に、従来用いられてきた各種の分析手法に対してこれらの情報を適用させられるようにするためには、それらを決まった長さ(個数)を持つ要素の組、すなわちベクトルで表す必要がある。メタ情報の方は、いわば一問一答もしくは一問多答

の形式の情報であり、ベクトル形式への変換は図１のように比較的自然に考えることができる。

公開年	IPC	出願人
2017	B23K 35/30	A社
2016	C22C 19/05|G05B 9/02|G06Q 50/06	B大学
2017	B23K 35/30|G05B 9/02	C研究所

2016	2017	B23K 35/30	C22C 19/05	G05B 9/02	G06Q 50/06	A社	B大学	C研究所
0	1	1	0	0	0	1	0	0
1	0	0	1	1	1	0	1	0
0	1	1	0	1	0	0	0	1

図１　属性情報のベクトル化

一方、テキスト情報のベクトル化については、筆者らも[4]で述べているが、テキスト内の単語の有無に基づいて（0,1）を要素とするベクトルで表現する方法(Bag of Words）や、ある単語の周辺状況を反映させるようにニューラルネットワーク手法で学習を行い、そこで得られた単語のベクトル表現（分散表現）を利用してテキストのベクトル表現を得る方法（word2vec をはじめとする Word Embedding 手法[5]+ SWEM[6]）が代表的なものとしてある。（図２）

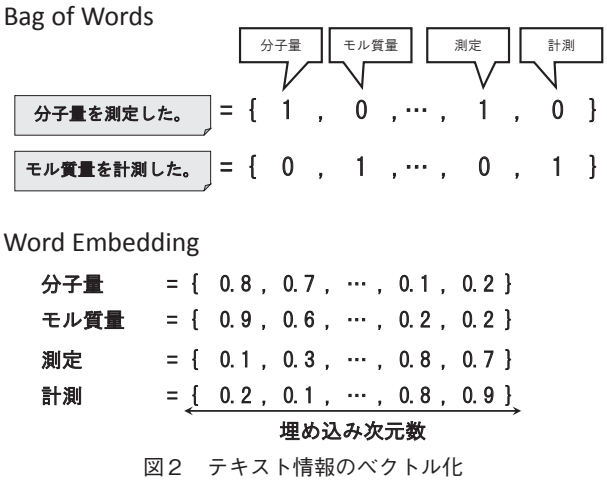

図２　テキスト情報のベクトル化

このようにして別々に得られたテキスト情報とメタ情報のベクトルを、単に連結してテキスト情報とメタ情報とを含めた文書全体のベクトル表現と考えることも可能である。この文書全体のベクトル表現を図３のように利用して分類モデルを構築することも可能であるが、Bag of Words のように単語の有無に基づいたテキストのベクトル化を行う場合、テキスト部分のベクトル次元数は利用するテキスト内の語彙数に等しくなるため、テキスト部分と属性部分とでどうしてもアンバランスが生じる。頻出単語や重要単語のみを利用してベクトル化する、ノイズになりうる単語を省く、など情報量を減らすことも考えられるが、その基準はどうあるべきか、別途その検証が必要になってくる。

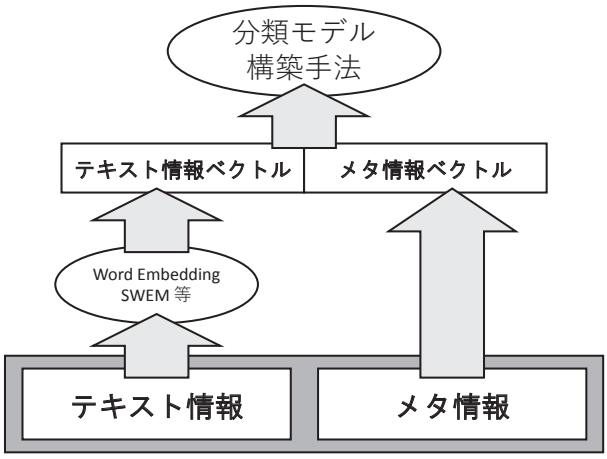

図３　ベクトル化されたテキスト・メタ情報の利用

また、テキストのベクトル表現として分散表現を利用する場合は、ベクトル次元数はモデル構築者が指定するため上記で述べたようなアンバランスさが発生する懸念は解消されるが、この分散表現そのものが果たして妥当なものであるのか、またテキスト間の類似度を適切に表現できているものかどうかこちらもまた検証が必要である。さらに、この学習による分散表現獲得の過程は「文書を分類する」というタスクとは独立に、無関係に行われるものであることは認識しておきたい。この過程には、必ずしも「文書を分類する」ことにおいてふさわしい分散表現を積極的に得ようとする働きは含まれていないということである。

2.2　テキスト情報とメタ情報を同時に扱う分類モデル構築手法

前節で、メタ情報のベクトルとテキスト情報のベクトルとをそれぞれ独立に獲得してそれらを結合する際の懸念点について述べたが、テキスト情報を扱うに当たってはそれらが１件１件異なる長さを持ち、さらに要素（単語）の並び・順番に意味があるという前提を事前に受け入れて学習を行うニューラルネットワークのモデルも広く知られており、筆者が[7]で利用した RNN（Recurrent Neural Network）はその１つである。

ここで、テキスト由来の情報は RNN で扱い、属性由来のメタ情報は通常のフォードフォワード型ニューラルネットワーク（多層パーセプトロン, MLP）で扱い、それらを結合した情報を利用して更に分類を行うようなネットワークをここでは提案する。（図４）

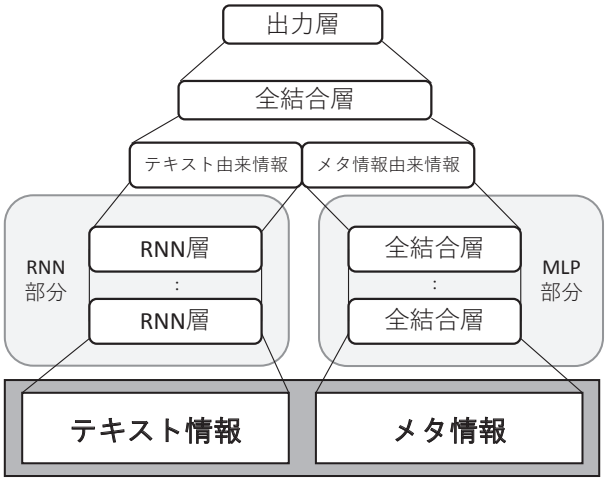

図４　テキスト・メタ情報を同時に扱う分類モデル

１件の文書がこのネットワークが入力されると、RNN 部分にはテキストの単語が順繰りに与えられてその内部状態が変化していき、最終的に RNN 部分の出力情報が得られる。別途、MLP 部分には属性情報が与えられ、ニューラルネットワークのフィードフォワード演算を経て出力情報が得られる。この両者の出力情報（ベクトル）を結合し、このテキストと属性とを結合した情報に対して更に演算を行って出力値が得られるが、この出力値と本来のあるべき予測値（教師データ）との誤差に基づいて、RNN 部分と MLP 部分全体のパラメータを調整していく。その結果、テキスト由来の情報を扱う部分も、メタ情報由来の情報を扱う部分も、正しい分類結果を導き出すという目的のもとで全体が最適化され、互いの情報を有効に利用して学習を行えることが期待できる。

3 特許文書分類モデルへの応用

本節では、注目すべき特許に人が手動でラベルを付与してその状況を学習し、さらにその他の特許に対して注目に値する特許を自動的に抽出する、という問題設定を想定して 2.2 章の手法を試行する。

データ概要や問題設定を表１に示す。

表１　データ概要と問題設定

対象データ	原子力関連分野の国内特許公報 500 件 うち注目特許 123 件、 非注目特許 377 件
学習データ	データ全体の 80%（400 件） うち 注目特許 99 件、 非注目特許 301 件
検証データ	データ全体の 20%（100 件） うち 注目特許 24 件、 非注目特許 76 件
テキスト情報	「要約」部分
メタ情報	テーマコード、 学習データ 400 件内で３件以上の 文書に出現したもの（全 56 種類）
教師データ	注目特許か否か、の２値カテゴリ情報

注目特許か否か、という教師情報は、客観的かつ正解との照合を可能にすることを考え、ここでは特定の IPC 分類を含むか否か、によって機械的に与えた。また、2.2 章の手法との比較対象のために、メタ情報部分は用いずテキスト部分のみを RNN を用いて分類するモデルを同一の問題設定のもとで別途構築した（図５）。

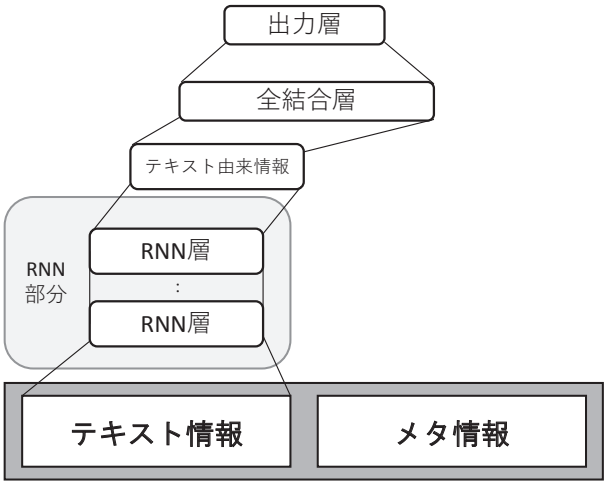

図５　比較対象：テキスト情報のみ（RNN のみ）

以下、結果を示す。「注目特許」クラスに関しての値であり、「RNN+MLP」が 2.2 章で今回提案した手法、「RNN のみ」が比較対象であるテキスト部分のみを用いて分類を行う手法である。

表２　注目特許分類 結果 精度情報

	F-score	precision	recall
RNN+MLP	**0.80**	**0.86**	0.75
RNN のみ	0.79	0.79	**0.79**

本試行の結果では、予測結果に対する正解率（precision）は RNN+MLP モデルの方が良好であり、元来の注目特許がどれだけカバーできたか、という指標（recall）は RNN モデルの方が良好であった。これに

より、RNN+MLP モデルの方が若干取りこぼしは多くなっているものの「より確実」なものを拾い上げているといえ、テキスト情報のみでは注目特許の可能性が高いが、メタ情報をあわせて考えると注目特許の可能性が低くなるようなものが存在していたと推測できる。正解率を再現率の調和平均である F-score では、若干であるが RNN+MLP モデルの方が良好であった。

4 まとめ

テキスト情報とメタ情報とを同時に利用して分類モデルを構築することで、双方を活用した文書分類が行える可能性を示した。用いる属性情報の拡張などを行い、さらなる試行・実験を通して今後は適用の場面を拡張していきたいと考えている。

当社(株)NTT データ数理システムの取り組みとして、利用者自身が持つテキストデータに対して言語解析処理を行って単語の抽出を行い、またそれらに対して RNN をはじめとする深層学習のアルゴリズムを適用させられるツールを開発・販売している。これらは共通の基盤プラットフォーム上で提供され、ツール間をシームレスに連係させて利用することができる。図6に、テキストマイニングツール Text Mining Studio の解析結果に対して、深層学習ツール Deep Learner を適用させてテキスト情報・メタ情報の双方を利用した分類モデルの学習を行っている様子を示した。これら分析ツールの適用により、多様化するデータに対しての様々な要請に応えることができると当社では考えている。

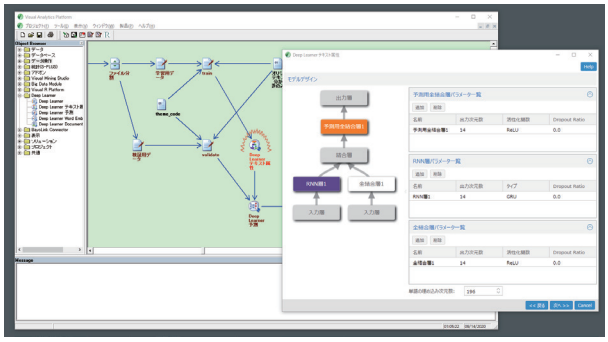

図6　Deep Learner の利用

参考文献

[1] 富永泰規：“外国特許文献へ分類付与に関する機械学習活用可能性調査について”Japio YEAR BOOK 2017 pp.212-217 (2017)

[2] 安藤俊幸：“機械学習を用いた効率的な特許調査法”Japio YEAR BOOK 2019 pp.240-251 (2019)

[3] 安藤俊幸：“機械学習を用いた効率的な特許調査法”Japio YEAR BOOK 2016 pp.150-161 (2016)

[4] 岩本圭介，柿沼匡志：“一般語彙と専門語彙の Word Embedding を併用した特許文書間の類似度算出”Japio YEAR BOOK 2019 pp.252-257 (2019)

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111-3119.

[6] D. Shen et al., “Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 2018, pp. 440-450.

[7] 岩本圭介：“ディープラーニングによる特許文献からの技術用語抽出”Japio YEAR BOOK 2017 pp.242-247