E-005

機械翻訳を活用した特許公報分類支援システムの提案
A Study of a Patent Classification System with Machine Translation

樗松理樹 ‡
Masaki Kurematsu

1. はじめに

企業等においては，代表的な知的財産情報である特許公報[1]の処理の 1 つとして，分類がある．特許公報には，IPC や FI 等の分類コードが付与されているが，請求内容に伴う面，申請者が独自に付けているため，分類者との観点の違いが生じるなどの課題がある．この問題に対し，独自の分類を特許候補に付与することが考えられる．しかし，人手で分類を付与することは，特許の種類・量などの点から作業負荷が大きいため、その作業を支援する手法が求められている。著者は、本課題を解決するために、企業の担当者の協力のもと，手法の構築[2]を行っている。これまでの提案手法は，基本的に専門家が分類を付与した特許とその特許中の語句の出現頻度との相関から分類器を構築することを試みている。これまでにナイーブベイズやラフ集合理論を用いた手法を提案してきたが，特許中に生じる語句の表記ゆれへの対応が大きな課題となっている。

この課題を解決するためには、表記ゆれがある語句を一定の語句に統一することが考えられる。この点に対し、本研究では、技術進歩とともに精度や利用しやすさが向上している機械翻訳[3]に着目する。機械翻訳を用い、英語に変換することで，表記ゆれのある語句を同一語句にまとめることが期待できる。また誤った翻訳（以後，誤訳）があるとしても、同じ機械翻訳器を用いれば、同様の誤訳となれば、同様に表記ゆれを吸収できると考えられる。

以上の背景から，本研究では，機械翻訳を利用した特許公報分類支援システムを提案する。本システムでは，はじめに分類済み特許公報の文章を機械翻訳で英文に翻訳する。その英文から構築した文書語彙行列（以後、DTM）に対し，ナイーブベイズ分類（以後、NBC）、K-NN、SVM を適用し、分類モデルを生成する。この分類モデルを用いて、新規特許の分類を試みる。以後、提案手法の詳細を述べたあと、評価実験の結果について述べる。

2. 機械翻訳を活用した特許公報分類支援システム

2.1 提案手法

図 1 に提案手法の概要を示す。本手法では，はじめに特許から抽出した文を機械翻訳で英文に翻訳する。さらに英文を品詞解析し，名詞を抽出する。次に，それらの語句から構築した DTM から分類モデルを構築する。この分類モデルを用いて，新規特許公報の分類を行う。以下，各手順の詳細を示す。

2.2 対象とする特許公報

本研究では，専門家によって一定の範囲に絞り込まれた特許公報を対象とする。これらの特許公報に対し，専門家は，特許が解決しようとする課題と課題を解決するための手段について，それぞれの分類を示す課題分類ラベル，手段分類ラベルを付与する。課題分類ラベルと手段分類ラベルは，大分類１つと小分類１つから構成される。また，大分類ごとに小分類はことなる。本研究では，作業者による揺らぎの大きい小分類は対象とせず，大分類のみを対象とする。

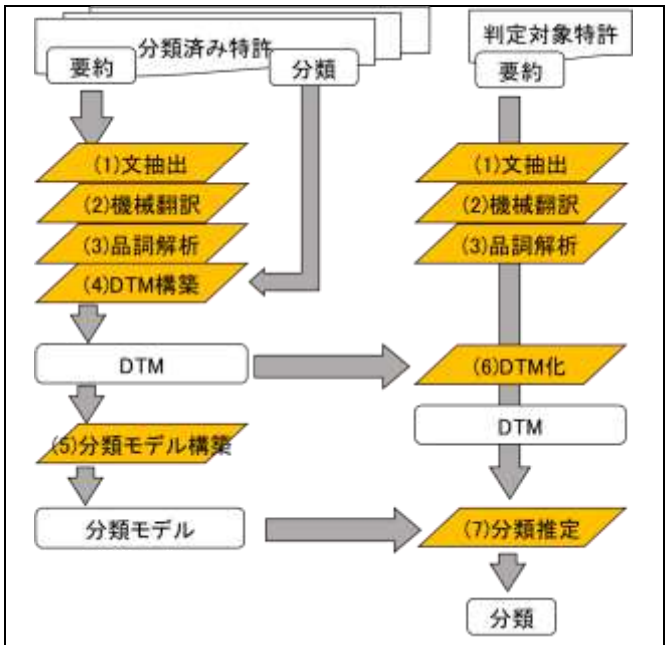

図 1　システム概要

2.3 対象とする文の抽出

本手法では、はじめに特許公報から処理対象となる文を抽出する。対象とする文としては、特許公報に含まれる要約文に着目する。これは、特許全文を扱う場合、ノイズとなる文が混ざると考えられるためである。要約文は，“【課題】文 $_1$、…，文 $_n$【解決手段】文 $_{n+1}$，…，文 $_m$”または“【目的】文 $_1$，…，文 $_n$【構成】文 $_{n+1}$，…，文 $_m$”のような構造をとる．このうち，文 $_1$ から文 $_n$ を課題について述べている課題文，文 $_{n+1}$ から文 $_m$ を手段について述べている手段文として抽出する．

上記で得られた文に対し、機械翻訳を行う。機械翻訳の結果、翻訳が出来なかった語を含む文を除いたものを、以降の処理の対象とする。

2.4 DTM 構築

得られた特許翻訳文に対し，品詞解析を行い，名詞のみを取り出す。さらに，これらの名詞のうち，出現回数が 2 回以上の名詞のみを抽出する。各文において，抽出された名詞の出現の有無から DTM を作成する。DTM は，課題，手段それぞれに作成する。また，このとき、Stemming 処理，StopWord 除去も行う。

2.5 分類モデル構築

2.4 で構築した DTM を元に，分類モデルを構築する。分類に利用するクラスとしては、特許に付与された分類を利用する。分類モデルとしては，分類が付与されている点，分類が目的である点から，教師あり機械学習手法を用いる。

2.6 判定対象特許の分類

分類判定の対象となる特許に対し，2.3 で述べた方法で，対象となる文の抽出，機械翻訳を行う。さらに，2.4 で構築した DTM に含まれている単語に従い，判定対象特許に対する DTM を構築する。得られた DTM に対し，2.5 で構築した分類モデルを適用し，分類を推定する。

3. 評価実験

3.1 実験概要

本提案手法の有用性を検証するために、次に示す評価実験を行った。

(1)評価データ：評価データとしては、専門家から提供をうけた分類済み特許公報を， 1998 年以前の特許 297 件（Data-1），1998 年から 2008 年まで特許 283 件（Data-2），2009 年から 2010 年の特許 59 件（Data-3）に分割し， Data-1，Data-2，Data-1 および Data-2 から構築した分類モデルを用いて Data-3 の分類を推定する．それぞれ，事前に付けられた分類を正解とし，抽出結果における順位との比較により評価する．

(2)機械翻訳：機械翻訳としては、株式会社クロスランゲージ社製の PAT-Transer V12 for Windows[5]を用いた。このツールは、特許明細書専用と銘打っており、本研究との相性がよいと考えられるため、選択した。

(3)品詞解析：英語の品詞解析には、Helmut Schmid 氏が開発した TreeTagger[4]を用いた。

(4)分類モデル：分類モデルとしては、代表的な教師あり学習モデルである、単純な確率的分類器である NBC、特徴空間における最も近い訓練例に基づいた分類の手法であり、パターン認識でよく使われる K-NN、教師あり学習を用いるパターン認識モデルの 1 つであり、分類や回帰へ適用できる SVM を用いる。これらの手法は、SPAM の分類[5]などに利用されていることから、本研究でも利用する。なお、実行には R 言語[6]のライブラリを利用している。また各手法におけるパラメータとしては、NBC におけるラプラス係数は 1、K-NN における K の値は 3、SVM におけるカーネル関数は、rbfdot を用いた。

(5)日本語による処理：比較対象として，日本語のままでも同様の処理を行う。日本語においては，翻訳が行われた文に対し，形態素解析を行う。得られた形態素列において，(a)形態素列，(b)2 文字以上のカタカナ列，(c) 2 文字以上の英字列の方法で語句を抽出する。ここで形態素列とは，名詞，語尾，形容動詞語幹が連続する部分を意味する。

3.2 実験結果

実験結果を表 1、表 2 に示す。表において正解率は，テスト用データ中分類が正しく判定された割合を示す。また，KNN，SVM において TF は出現回数を，B は 0,1 に変換した結果を利用したことを示す。

表 1　実験結果（課題：正解率）

データ	D1+2		D1		D2	
言語	英	日	英	日	英	日
NBC	26	34	53	32	55	26
KNN(TF)	53	9	21	11	26	12
KNN(B)	28	11	22	13	22	9
SVM(TF)	29	15	17	9	28	15
SVM(B)	26	15	16	9	26	15

表 2　実験結果（手段：正解率）

データ	D1+2		D1		D2	
言語	英	日	英	日	英	日
NBC	55	26	78	41	74	34
KNN(TF)	26	12	45	13	43	10
KNN(B)	22	9	43	14	34	11
SVM(TF)	28	15	60	15	40	15
SVM(B)	26	15	55	15	41	15

3.3 評価・考察

実験の結果から，機械翻訳結果を用いた場合が，正解率が向上している。これは，単語の絞り込みがうまくいったためと考えられる。また 3 種の分類モデル NBC のほうが正解率は高い。また，利用する文数が多い場合のほうが，正答率が低くなる傾向がある．これはモデルに含まれる語句と新規特許中の語句とのずれが発生しているためと考えられる。この点から，機械翻訳により語句の揺れはある程度は抑えられるが，まだ十分ではないと考えられる。

今後の課題としては、結果の精査，英語の類義語辞書を用いた統合，NBC などとは異なる手法の適用による精度向上，および同一データによる評価があげられる。

4. おわりに

本稿では，特許公報処理支援を行うために，特許公報で述べられている，解決しようとする課題とその手段について、過去の事例から求めた分類モデルを用いて分類する手法を提案した。本手法の大きな特徴としては、語句の表記ゆれを解決するために、機械翻訳で一度英文に翻訳したあと、分類モデルを構築する点にある。専門家が分類付けした特許公報を利用した実験の結果、日本語を用いた場合よりも精度の向上が見られた。このことから、本提案手法の有用性が示せた。今後の課題としては、単語の抽出方法の改善、分類モデル作成方法の検討、再評価があげられる。

謝辞

評価実験にご協力いただいた A 氏に感謝の意を表します。また本研究の一部は，科研費・基盤 C（課題番号 15K00154）の助成を受けております。

参考文献

[1] 社団法人発明協会，“産業財産権標準テキスト 特別編”，東京書籍（2005）
[2] 樽松理樹, “タイトル”, 論文誌名, Vol.n, No.n (2018).
[3] クロスランゲージ社，PAT-Transer V12 for Windows，https://www.crosslanguage.co.jp/products/pat-transer_v12/
[4] Helmut Schmid，TreeTagger，http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[5] Brett Lantz，“Machine Learning with R”，Packt Publishing，(2013)
[6] R-Project，https://www.r-project.org/