The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

機械翻訳とラフセット理論を利用した特許公報分類システム

A Study of Patent Publication Classification Using Machine Translation and Rough Set Theory

樽松理樹[*1]
Masaki KUREMATSU

[*1] 岩手県立大学
Iwate Prefectural University

Abstract: It is important to check exists patents before submitting own patents or sailing new products. However, it is hard task to check a lot of patents. In order to support this task, I proposed a framework of a patent publication classification system using machine translation and Rough set theory in this paper. It makes a classifier from patent publications labeled by experts with the following 4 steps. In step.1, this framework extracts sentences from abstracts of patents based on block tags. In step.2, it translates these sentences to English using Machine translation and extracts terms using Term Frequency and Rough Set reduction. In step.3, it makes a Document Term Matrix form extracted terms. In step.4, it makes a Naive Bayes Classifier and Rough set rules from a Document Term Matrix as classifier. It classifies unlabeled patent publications by these classifiers. I developed this framework by R language and some natural language processing tools and evaluated. In evaluation, I tried to classify some patent publications with an expert. Experimental results show the possibility of this approach.

1. はじめに

特許公報[発明協会 05]は，代表的な知的財産情報である．特許公報の処理として分類を行うことが多い．検索システム[藤井 12]を用いて行うことは可能であるが，実務においては，さらに細分化をする必要がある．研究協力者である企業の知的財産部門に所属する専門家は，その特許が述べている課題と手段で分類している．しかし，特許公報が膨大であるため，このような独自の処理に対応するツールが必要となっている．

この課題に対し，著者は企業と協力し，手法の構築 [樽松 18]に取り組んできた，これらの研究においては，人による検証・反映を意図し，理解できる分類モデルを構築することと目標に，表層情報を用いるナイーブベイズ分類（以後，NBC）を中心に進めてきた．しかし，分類精度は不十分である．その要因として，表記ゆれ，情報不足などが挙げられる．

以上の背景から本研究では，本研究では，機械翻訳とラフセット理論を利用した特許公報システムを提案する．本手法では，機械翻訳で英語に翻訳することで表記ゆれの削減を図る．また，ラフセット理論[Zdzislaw 82]における「属性の縮約」（クラスがお互いに識別されるために必要かつ十分な属性の組）を用いることで語句の絞込みを，決定ルールを用いることで語句間の関係を考慮することを図る．

2. 提案システム

2.1 システム概要

本システムは図1に示すように，大きく「DTM 構築部」「分類器作成部」「分類推定部」からなる．DTM（Document Term Matrix）とは，各列が語句，各列が文書に対応し，各文書における語句の出現数を行列形式で示したものである．「DTM 構築部」では，専門家によって与えられた特許公報から，分類処理に有用と評価し抽出した語句の文書毎の出現数から DTM を構築する．「分類器生成部」では，ラフセット理論に基づき，DTM から決定ルールを抽出する．また、NBC で利用する尤度表を構築する．「分類推定部」では，決定ルールおよび尤度表をもとに新たな特許の分類を行う．以降で各部分の説明を加える．

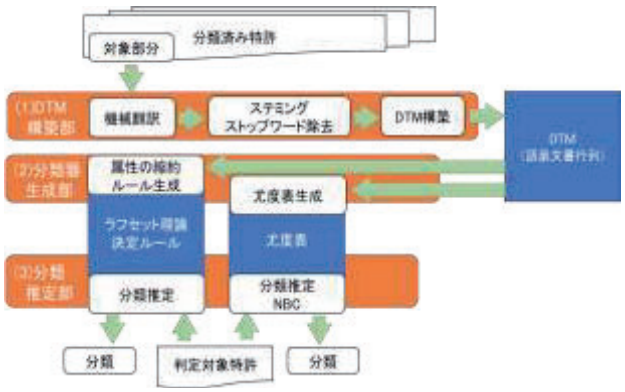

図 1. システム概要

2.2 対象とする特許公報

本システムでは，専門家によって ICP や F タームなどで絞り込まれた特許公報を対象とする．これらの特許公報から、独自の分類を行うために着目する部分を，ブロックタグを利用して抽出する．ブロックタグとは，特許に含まれるブックマークであり，特許の構成要素を示している．これを利用することで，特許から必要な部分が抽出可能となる．以後，抽出された部分を本稿では特許文と呼ぶ．

2.3 DTM 構築部

専門家に分類付けされた特許公報から，以下の流れで分類出現語句情報を抽出する．

① 機械翻訳を用いて特許文を英語に翻訳する．
② ①で得られた英文から，事前に指定する品詞の単語のみを取り出す．なお複数回出現する場合も 1 回のみ取り出す．取り出した単語に対し，ステミング，ストップワードの削除を行う．
③ ②で抽出した語句を元に，DTM を構築する．このとき，DTM の各要素は，出現の有無を示す 2 値とする．

連絡先：樽松理樹，岩手県立大学ソフトウェア情報学部，岩手県滝沢市巣子 152-52，電話：019-694-2582，FAX：019-694-2501，メール：kure@iwate-pu.ac.jp

2.4 分類器生成部

前述までの処理で構築した DTM から，(1)ラフセット理論に基づく決定ルール及び(2)NBC に必要となる尤度表をそれぞれ構築する．決定ルールを構築するために，はじめに属性の縮約を行う，その後，決定ルールを構築する．尤度表とは，各語句について，特定の分類の文書において出現した割合を示す表である．これは単純に数え上げることで作成する．

2.5 分類推定部

分類推定部では，推定対象となる特許に対し，ラフセット理論に基づく決定ルールおよび尤度表を用いた NBC を適用する．それぞれ最も評価値が高い分類を正解として推定する．

3. 検証評価

3.1 検証内容

提案手法の有用性を評価するために，専門家の協力のもと検証を行った．以下段階を追って説明する。

(1) 対象とする特許公報：専門家が分類わけを行った特許公報 470 件を用いる．各特許公報は，6 種類に分類わけされている．分類ごとの特許数は最大 118，最小 47，平均 78.3±27.6 と偏りがあるが．このまま利用した．

(2) 特許文抽出：今回利用する分類は，特許公報に挙げられた「解決すべき課題」である．その点から，課題に関する記述部分を対象とする．また明細書を利用する場合，ノイズが含まれることが多くなると考えられるため要約中の課題または目的に関する記述の部分を用いる．

(3) 機械翻訳：今回は，特許専用と銘打っているクロスランゲージ社製の PAT-Transer V12 for Windows[クロスランゲージ 14]を用いて翻訳を行った．

(4) 語句の抽出・DTM 構築：機械翻訳で得た翻訳文に対し，Helmut Schmid 氏が開発した TreeTagger[Schmid 94]を用い，名詞，動詞の動名詞・現在分詞・過去分詞を対象とした．これらは専門家との話し合いで決定している．さらに R 言語[The R Foundation]の tm パッケージ[Feinerer 18]を利用し，ステミング，ストップワードの削除を行う結果から，DTM を構築する．

(5)分類器生成：構築した DTM に対し，ラフセット理論に基づく決定ルールと尤度表の作成を行う．これらには，R 言語の RoughSets パッケージ[Riza 15]，e1071 パッケージ[Meyer 19]をそれぞれ用いた．なお NBC において Laplace は 1 に設定している．

(6) 予測：予測では，5-fold Cross Validation（訓練データ：平均 378 文，テストデータ：平均 92 文）を実施する．この際ラフセット決定ルールの投票は，成立したルールの Laplace 評価値の最大値を用いた．

3.2 検証結果

評価においては，Accuracy および Kappa(K 統計量)を用いる．Accuracy は正しく分類を推測した割合・精度であり，Kappa は偶然正しい予想になった確率を差し引いて Accuracy を調整した値である．また，テストデータ中の 1 クラスの割合の最大値 No Information Rate（NIR）は 0.19 である．

本検証においては，NBC では 7 文書以上出現する語句を用いた．これは出現数の平均が 7.7 であったことから決定した．結果，平均 150 個の語句を抽出した．一方，ラフセット理論による属性の縮約では平均 50 個の語句を抽出した．

分類の推定結果を図 2 に示す．図 2 において#は Cross Validation の回数，RST はラフセット理論の結果，ランダムは，ランダムで選択した場合の正解確率（1/分類数）を意味する．

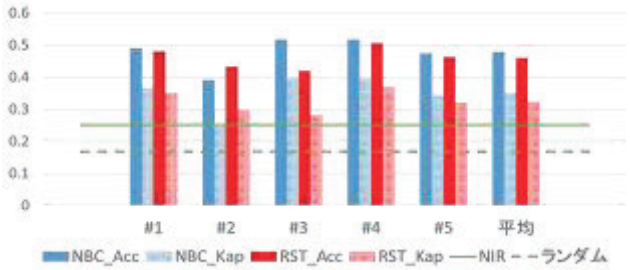

図 2 検証結果

3.3 評価

NBC，ラフセット理論による分類ともには，Accuracy，Kappa ともに NIR を上回ったことから，一定の効果があると考えられる．また NB とラフセット理論による分類の間には有意差があり，NBC の精度の方が高い．

4. おわりに

本稿では，機械翻訳とラフセット理論を用いた特許分類推定手法を提案した．評価実験の結果，NBC を用いた場合は一定の効果が見込まれたが，ラフセット理論を用いた結果は十分な成果を得られなかった．今後の課題としては，設定を変更した検証の実施を含めた検証結果の分析，その結果に基づく手法の改善，他の文書分類課題での適用結果を踏まえた手法の長所短所を明らかにすることが挙げられる．

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

[クロスランゲージ 14] クロスランゲージ社，PAT-Transer，https://www.crosslanguage.co.jp/products/pat-transer_v12/ ,2014

[Feinerer 18] Ingo Feinerer，Kurt Hornik，Artifex Software, Inc.The：Package ‘tm’，https://cran.r-project.org/web/packages/tm/ , 2018

[藤井 12] 藤井敦，谷川英和，岩山真，難波英嗣，山本幹夫，内山将夫：特許情報処理:言語処理的アプローチ，コロナ社，2012

[発明協会 05]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

[樽松 18] 樽松理樹：ラフセット理論を用いた特許公報分類支援システムの提案，人工知能学会全国大会第 32 回，2018

[Meyer 19] David Meyer， et.al : Package ‘e1071’， https://cran.r-project.org/web/packages/e1071，2019

[Riza 15] Lala Septem Riza, et.al ：Package ‘RoughSets’，https://cran.r-project.org/web/packages/RoughSets/，2015

[Schmid 94] Helmut Schmid: Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing，pp.44-9 ，1994 ，http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[The R Foundation 97] The R Foundation，https://www.r-project.org/，1997

[Zdzislaw 82] Pawlak，Zdzisław :Rough sets，International Journal of Parallel Programming，Vol.11，No.5，pp.341–356，Springer，Heidelberg , 1982