E-011

化学分野における特許発明の効果を予測する深層学習モデル
Predicting the effects of chemical patent inventions with deep neural networks

高橋 林太郎† 正田 備也‡
Rintaro Takahashi Tomonari Masada

1. はじめに

近年，多くの産業分野において機械学習の利用が検討されているが，この流れは特許などを取り扱う知財業界にも波及している．特許を取得するためには，特許庁に特許出願を行い，審査を受ける必要がある．ここで特許出願において，最も重要なセクションは「特許請求の範囲」である．主たる審査対象であり，また審査を経て特許が登録された後には特許権の権利範囲を定めるものとなるからである．そして特許出願の審査においては，特許請求の範囲に記載された発明が奏する効果，即ち「発明の効果」が考慮される．発明が従来技術から予測できない効果を奏する場合，特許性が肯定され易くなる．特に化学分野の発明は，機械分野やソフトウェア分野などの発明と異なり，特許請求の範囲に記載された発明の構成から当該発明が奏する効果を予測することは困難である．なぜならば，化学反応や化学物質同士の相互作用は必ずしも人間の予想通りにはいかないことが多く，実際に試験をしてみなければ分からないからである．

本稿では，発明の効果を予測することが難しい特許の例として，「重合体」に関する特許に注目する．そして，「重合体」及びその類義語を特許請求の範囲に含む特許発明について，当該発明が「接着性」に関する効果を発現するものか，又は「接着性」とは一見相容れない「剛直」といった効果を発現するものかを，特許請求の範囲のテキストから予測する深層学習モデルを提案する．

2. 従来研究

特許請求の範囲を解析する手法が提案されている．新森らは，特許請求の範囲での特有の表現に着目し，特許請求の範囲の可読性を向上させる手法を提案した[1]．Suzuki らは，請求項から新規性に関するキーワードを抽出する手法を提案した[2]．前原は，特許請求の範囲における上位概念化された記載に対応する，明細書中の下位概念の文章を抽出する手法を提案した[3]．そして西尾らは，インクジェットメディアの技術分野における特許請求の範囲を入力文とした教師あり機械学習で文書分類を行うに際し，機械学習モデルに入力する文書ベクトルの違いが精度に及ぼす影響について報告した[4].これらの研究とは異なり，本稿では発明の効果を予測するために特許請求の範囲を解析する．

3. データ

3.1 「特許請求の範囲」と「要約」

本稿に用いるデータを取得するに際し，審査を経て登録された特許について，その「特許請求の範囲」と「要約」を検索対象とした．一例として，特許第 6891366 号の「特許請求の範囲」の一部と，「要約」とを抜粋して図１に掲載する．

【特許請求の範囲】
【請求項１】（Ａ）エーテル基を有するポリウレタン樹脂９０～７０重量%、及び、（Ｂ）ポリオキシアルキレン基及び炭素数１～３０の炭化水素基を有する単官能エポキシ化合物１０～３０重量%を含有する親水性コーティング用樹脂組成物。
【請求項２】単官能エポキシ化合物（Ｂ）のポリオキシアルキレン基が、ポリオキシエチレン基である請求項１の親水性コーティング用樹脂組成物。・・・

【要約】
【課題】液安定性が高く、耐水性や密着性を有するとともに親水性に優れる塗膜を形成できる親水性コーティング用樹脂組成物を提供する。
【解決手段】（Ａ）エーテル基を有するポリウレタン樹脂９０～７０重量%、及び、（Ｂ）ポリオキシアルキレン基及び炭素数１～３０の炭化水素基を有する単官能エポキシ化合物１０～３０重量%を含有する親水性コーティング用樹脂組成物に関する。
【選択図】なし

図１ 「特許請求の範囲」と「要約」の一例

3.2 検索

検索には，特許検索・特許分析サービス「パテント・インテグレーション」を使用した．「特許請求の範囲」と「要約」を検索対象とした検索を行い，「接着性」等に関する特許を「ソフトクラス」とし、「剛直」等に関する特許を「ハードクラス」としてデータを収集した．

データ数を確保するため，「重合体」,「接着性」,「剛直」に加えてそれらに類似する用語も検索ワードに含めた．また特許請求の範囲，即ち発明の構成自体に「接着性」，「剛直」など効果に直結するキーワードを含む特許は除外した．単純なキーワードのマッチングのみで発明の効果を予測しうる特許を除外するためである．そして一方のクラスを検索するに際し，他方のクラスについてのキーワードを含む特許は除外した．それぞれのクラスを検索する際の条件を以下に示す．

3.2.1 ソフトクラス

・条件 1：「接着性」「粘着性」「密着性」の少なくとも何れかを【要約】に含むこと．
・条件 2：「重合体」「ポリマー」「重合物」「ゴム」「樹脂」「エラストマー」の少なくとも何れかを【特許請求の範囲】に含むこと．

† 立教大学 Rikkyo University
‡ 立教大学 Rikkyo University

・条件 3：「接着」「粘着」「密着」の全てを【特許請求の範囲】に含まないこと．
・条件 4：「剛直」「機械的強度」の全てを【要約】と【特許請求の範囲】に含まないこと．

3.2.2 ハードクラス

·条件 1：「剛直」と「機械的強度」の少なくとも何れかを【要約】に含むこと．
·条件 2：「重合体」「ポリマー」「重合物」「ゴム」「樹脂」「エラストマー」の少なくとも何れかを【特許請求の範囲】に含むこと．
·条件 3：「剛直」及び「機械的強度」を【特許請求の範囲】に含まないこと．
·条件 4：「接着性」「粘着性」「密着性」の全てを【要約】と【特許請求の範囲】に含まないこと．

3.3 学習データとテストデータ

それぞれクラスについて 4,644 件のデータを取得した．4,644 件のデータを公開日順に並べ，新しいデータの側から 500 件ずつ，合計 1,000 件のテストデータを切り出した．新しいデータの側からテストデータを切り出した理由は，既存の発明を元に未知の発明の効果を予測するという，本稿のモデルについて想定される使用態様に合わせるためである．テストデータを切り出した残り，即ちそれぞれのクラスについて 4,144 件，合計 8,288 件を学習データとした．

全てのデータについて「特許請求の範囲」のデータクレンジングを行い，「請求項 1」のテキストのみを取得した．

4. 実験

本稿では Support Vector Machine を用いた分類モデルと，Long Short-Term Memory[5]を用いた分類モデルの 2 種類をそれぞれ学習させた．

以下では，Support Vector Machine を「SVM」と称し，Long Short-Term Memory を「LSTM」と称する．

4.1 SVM を用いた分類モデル

8,288 件の学習データの請求項 1 を MeCab で形態素に分割した．形態素に分割した請求項 1 を，Python の scikit-learn ライブラリが備える TfidVectorizer でベクトル化し，説明変数とした．語彙数は 4,687 であった．入力データである説明変数に対応させ，ソフトクラスを 0，ハードクラスを 1 とするラベルを作成し，目的変数とした．

学習データを 5 つに分割し，そのうちの 1 つを検証データとする交差検証を行い，2 クラス分類の学習を行った．

4.2 LSTM を用いた分類モデル

8, 288 件の学習データの請求項 1 を MeCab で形態素に分割した．形態素に分割した請求項 1 を用い，Keras ライブラリが備える Tokenizer クラスの fits_on_tests メソッドにより形態素ごとにインデックス番号を割り当てた．次いで，形態素に分割した請求項 1 を texts_to_sequences メソッドに渡し，請求項 1 をインデックス番号の配列に変換した．そして，pad_sequences メソッドを用いて，所定の形態素数に満たない配列は 0 でパディングして配列の長さを一致させた．配列の長さを一致させたデータを説明変数とした．なお配列の長さ，即ち形態素の数は 300 とした．入力データである説明変数に対応させ，ソフトクラスを 0，ハードクラスを 1 とするラベルを作成し，目的変数とした．

ソフトクラスとハードクラスのそれぞれから 15%ずつデータを切り出し検証データとした．残りの 85%を学習に用い，Embedding レイヤー，3 つの LSTM レイヤー，2 つの Dropout レイヤー，Dense レイヤーを備えるモデルについて，2 クラス分類の学習を行った．

4.3 結果及び考察

以下の表 1 に、各モデルの分類性能の結果を示す。

分類モデル	正解率		
	学習データ	検証データ	テストデータ
SVM	−	0. 8404	0. 7400
LSTM	0. 9361	0. 8408	0. 7350

表 1　各モデルの分類性能

2 つのモデルのテストデータにおける正解率は同程度であった．一般的に LSTM は系列データの分類性能に優れると言われるが，請求項 1 のテキストの系列情報は，発明の効果の予測精度向上に大きくは寄与しない可能性が考えられる．また何れのモデルにおいても，テストデータにおける正解率は，検証データにおける正解率よりも低かった．これは新たな物質など新規技術を含むテストデータがモデルに入力された場合に，形態素が未知語として処理される等の理由で予測が困難になることが一因と考えられる．

5. おわりに

化学分野における発明の一例として，「重合体」等のキーワードを特許請求の範囲に含む特許発明を取り上げ，当該発明が奏する効果を 2 クラス分類で予測する手法を提案した．本稿の手法は，化学分野に精通した技術者のみならず，特許出願に携わる企業の知的財産部員，特許事務所の弁理士，そして特許庁の審査官が発明の効果を予測する際の一手法としての利用が期待できる．

本稿の学習及びテストに用いたデータには，モデルへ入力する請求項 1 に，テキスト以外に化学物質の構造や数式などの画像を含む特許も包含され，また特許請求の範囲全体には「重合体」等のキーワードを含む一方で，モデルへ入力する請求項 1 自体にはそれらを含まない特許も包含される．分類モデルのアーキテクチャやハイパーパラメータの調整のみならず，データの収集方針についても更なる検討を行うことで，本稿のモデルの利便性向上が期待できる．

参考文献

[1] 新森 昭宏, 奥村 学, 丸川 雄三, 岩山 真, “手がかり句を用いた特許請求項の構造解析”, 情報処理学会論文誌, Vol.45, No.3 (2004).
[2] Shoko Suzuki, Hiromichi Takatsuka, “Extraction of Keywords of Novelties from Patent Claims”, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (2016).
[3] 前原 義明, “トランスフォーマーを用いた特許審査支援の探究 - Detailed Description Is All We Need –”, 特技懇誌, Vol.297 (2020).
[4] 西尾 潤, 安藤 俊幸, “機械学習を用いた特許文書分類における入力ベクトルの影響”, 第 16 回情報プロフェッショナルシンポジウム予稿集 (2019).
[5] Sepp Hochreiter, Jurgen Schmidhuber, “Long Short-Term Memory”, Neural Computation, Vol.9, No.8 (1997).