特許情報収集による知的財産創造のための発見的価値創造の手法の開発

富山県立大学情報システム工学専攻　　1855005 小野田成晃

1 はじめに

ICT 分野の発達により, 民間団体や政府機関のデータを ICT 化することの重要性が増している. 総務省では, オープンデータ戦略の推進と題して, 行政の透明性・信頼性の往生, 国民参加, 官民協働の推進, 経済の活性化・行政の効率化が三位一体で進むことを目的として行われている [1]. 例えば, 災害関連情報として震源や震度に対するデータベースとその API を公開する試みが行われている [2]. そのうちの一つとして特許情報プラットフォーム[1]がある, そこでは日本の特許庁に提出された特許や実用新案等が掲載されており, Web サイト上で特許をキーワード検索することで特許利用の効率化を図っている.

しかし, これらのオープンデータは人手で少数の特許事例を調べるのには必要充分であるが, ビックデータとして特許全体の分析を行いたい場合は整理されているとはいい難い, 例えば, データの保存形式が PDF 担っている場合や, 被引用特許の件数が掲載されていない等の問題点がある.

特許文書ではテキスト情報の他に引用件数, 発明者, 出願年等の多数のデータが存在するそのデータをそれぞれ考慮しつつ, 新しい特許の組み合わせ等を提示すれば経営・開発の意思決定の一助となるであろう. そこで, 本研究はマルチモーダルなデータを利用するための特許生成支援モデルを提案する.

2 適用手法

2.1 言語生成モデルの概要

特許の生成のために自然言語処理分野で用いられている言語生成モデルを採用する. 言語生成モデルとは言語を生成するためのモデルとしてニューラル言語モデル（以下言語モデル）を利用するニューラルネットワークのことを指す.

言語生成するためには基本的に以下のプロセスで行われる

[1] 言語データから言語モデルを学習

[2] 学習したモデルを用いて単語・文を入力

[3] モデルによりその単語の次に尤も出現する単語を提示

言語モデルから文書を生成するイメージとしてを図 1 に示す. この例では”I” という単語が入力された際の次の出現単語の確率を再帰型ニューラルネットワーク (RNN) で出力してシークエンシャルに文書を生成することが可能である.

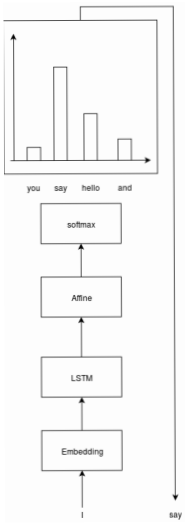

Fig. 1 生成モデルの概念図

2.2 LSTM

勾配消失・勾配爆発等の問題に対処するために, 言語モデルの RNN として LSTM を採用した. LSTM は以下の式で表される.

$$
\begin{aligned}
f &= \sigma(x_t W_x^f + h_t - 1 W_h^f + b^f) \\
g &= \tanh(x_t W_x^g + h_t - 1 W_h^g + b^g) \\
i &= \sigma(x_t W_x^i + h_t - 1 W_h^i + b^i) \\
o &= \sigma(x_t W_x^o + h_t - 1 W_h^o + b^o) \\
c_t &= f \odot c_t - 1 + g \odot i \\
h_t &= o \odot tanh(c_t)
\end{aligned}
$$

x: 入力データ, h: 隠れ状態, t: 時間, W: 層の重み, b: バイアス

2.3 言語生成モデルの理論

上記の RNN, LSTM を言語生成に適した形に応用したのが seq2seq という手法である. RNN, LSTM では異なる長さの入力に対応できなかった問題を解決した手法. sequence(系列) から sequence に変換する生成モデルを seq2seq と呼ぶ.

文章も系列データなので, 文章から文章への変換にも適用できる. seq2seq は言語情報を特徴マップに射影する Encoder 部分と特徴マップから言語に変換する Decoder 部からなる. 実用例としては以下の仏語から英語の翻訳ネットワークがある

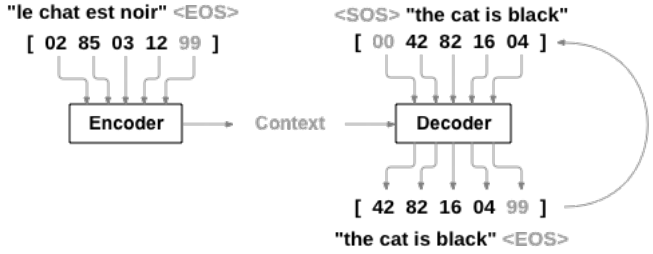

Fig. 2 フランス語から英語翻訳ネット

2.4 提案手法

前回生成される手法ではエンコーダーに LSTM を用いていた. しかし, このままでは特許の引用, 発明者等の被文書データを扱えない.

そこで, 画像キャプションで扱われるマルチモーダルモデルを参考に通常の seq2seq モデルのエンコーダー部分を多層 NN に変更することで特許の複雑なパラメータを考慮し且つ制御可能な特許生成モデルができると考えた.

エンコーダー部分に特許パラメータを入力とする多層 NN を適用する. そして特許の複雑なパラメータ情報を特徴マップとして出力してその圧縮された特許情報をデコーダーにかけることでパラメータを考慮した特許文生成が可能であると仮説をたてた. 図 3 は提案した言語生成器である. このように Encoder 部分に LSTM ではなく多層 NN 等の複数パラメータの重み付け・マッピングができる手法を選択することで入力を意思決定者の好みに合わせ, それに適した特許案を提案できるであろう.

エンコーダーをどのように改良するかが, 今回の特許提案データを作成するにあたっての肝となる.

従来のエンコーダ・デコーダモデルでは入力には一つの系列しか考慮されておらず, マルチモーダルに対応させる必要がある.

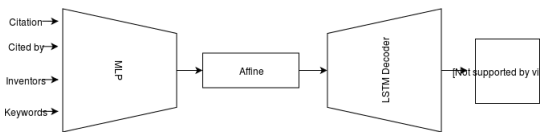

Fig. 3 提案したマルチモーダル言語生成器

[1]https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage

3 特許データ収集基盤

3.1 作成システムの概要

特許のオープンデータは人手で少数の特許事例を調べるのには必要充分であるが，ビックデータとして特許全体の分析を行いたい場合は整理されているとはいい難い．例えば, データの保存形式が PDF 形式の場合や, 被引用特許の件数が掲載されていない等の問題点がある．

そこで, 今回特許の定量的な分析をするためのリソースとして, 日本語ドメインの Patent - Google[2]がある．Patent - Google は Google 検索オプションの一つで, 世界各国の特許データが html 形式で公開されている．これは PDF などの非構造データに比べてデータ整理・収集しやすい利点がある．

またこの検索プラットフォームの他に Google Patent[3]がある．こちらは独自のドメインを持っており検索インタフェースと検索結果に多少の違いがある．現在の Google の特許検索の状況を整理するため図 2 を付す．Google Patent と Patent - Google はいずれも特許記事を patents.google.com ドメインで公開している．そのため 2 つの特許の文書に本質的差異はない．

本研究では, 検索オプションが豊富で通常の Google 検索エンジンと同様に使える Patent - Google を情報収集のプラットフォームとして利用した．

まず, 必要なデータとして,1. 特許 ID, 2. 発明者, 3. タイトル, 4. 承認日, 5. 引用特許数, 6. 被引用特許数, 7. 本文に含まれる特許内の単語とその頻度とした. また, 本研究では津村らと同様, 特許に含まれている単語を分析対象としているため, 文章から抽出する素性は名詞のみとした [?]. なお収集特許は実験後の単語を詳細に分析するため，著者の母語である日本語で提出されたものを対象とした．

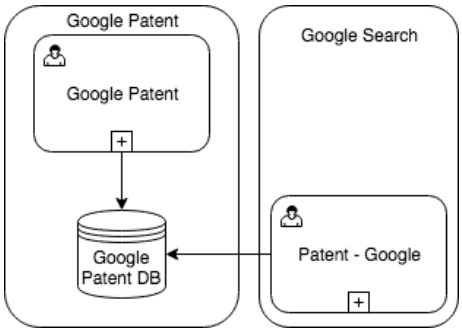

Fig. 2 Google における特許検索のプラットフォーム

3.2 データベース

Patent - Google から収集した特許データを蓄積・分析するためにデータベースを構築した．収集データは 7 種類あるが, そのうち単語に関しては各特許に対して抽出できる種類数が異なるため, スケーラビリティに富む NoSQL である mongoDB を用いた. また, 収集した全特許に含まれる全単語種を分析に用いるため, 別途全単語辞書を構築した．

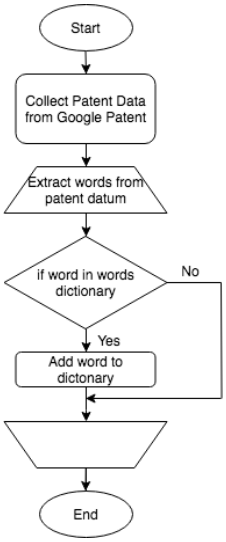

Fig. 3 単語辞書作成の手順

[2]https://www.google.co.jp/?tbm=pts
[3]https://patents.google.com

表 1: LSTM のパラメータ設定

使用モデル	Attention 付き Bi-LSTM
隠れユニット	50
損失関数	負の対数尤度
埋め込み次元数	50

次に収集した特許の全単語辞書を生成する手順を図 3 に示す．辞書には単語とその合計頻度を入れる．まず，前述 Google Patent を対象として各特許記事から上述の 7 個の情報を抽出・保存する．そこから特許データベースを作成する．次にそのデータベースに対して再帰的に単語データを検索しそのデータを辞書に加える．

そして, すべての含有単語を探索した後, 図 3 のループを終了して辞書の生成が完了する．なお, すでに単語が登録されていた場合は回数のみ加算する．

特許の価値付指標

[1] 被引用数：特許集合全体における被引用の数

[2] 平均被引用数：被引用数を公開年から現在までの経過年数で割った値

[3] エントロピー：2 つの特許の被引用が同じであれば, 公開から現在まで均等に引用されるロングセラーのが良いという仮説に基づいた尺度

[4] HITS：Kleinberg が提唱した web ページの重要度算出法．Google 検索エンジンにも利用されている．

3.3 教師あり学習のための特許内重要キーワードの抽出

アプリケーションのインタフェースとしてはユーザの関心のある特許キーワードを入力してもらう形を取るので, 教師あり学習でユーザの入力に対応するため, 学習用特許文書データから重要キーワードを抽出する必要がある．

そこで, 本研究では東京大学中川研究室と横浜国立大学森研究室が共同で開発した termextract という重要語抽出モジュールを用いた．このモジュールの大まかなプロセスとしては以下となる．

[1] 形態素解析プログラムによる単語分割

[2] 複合語の作成

[3] 文章中における重要度の計算

この 3 つのステップを経て, 複合語により複雑な概念を表すことが多い専門用語をキーワードとして文章中から抽出することが可能となった．

これを学習用特許文書に適応して一つの特許から任意の数のキーワードを取り出す．なお学習時は取り出したキーワードには重要度というスコアの降順に取り出す．

5 数値実験と考察

使用モデルは attention 付き LSTM を利用, パラメータ設定は表 1 のようにした．

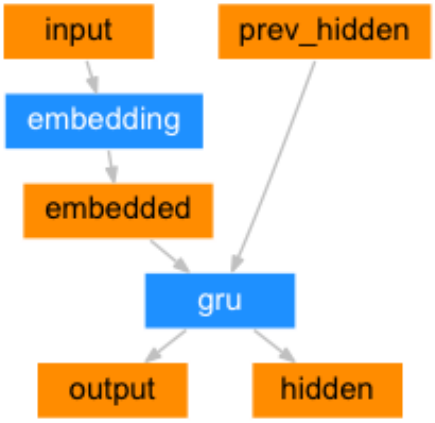

Fig. 3 使用する LSTM ネットワークの概念図

実験結果

学習結果の一部抜粋以下に学習結果の抜粋を載せる．学習結果 [1] では入力キーワードの文書とほぼ一致している文書ができている．ただし

表 2: 学習結果

学習時間	2944m 42s
エポック数	100
学習前の誤差	4025.8406
学習後の誤差	0.0430

後半にかけて文書の再現ができていないのでやはり長い文書では工夫が必要と考えられる.

学習結果 [2] では学習不足やその他の問題により power という単語の連続になっている. これは power という単語の次に来る単語の確率で尤も高いのが power となるという学習の結果と見られる.

このような例では数理最適化における局所解のように一度よいと思い込んだ予測結果から人間にとって間違いといえる結果を次々に出力する連鎖が起こっていることから**単語予測の沼**と名付けた.

[1] うまく行った例

入力 automation trainer solenoid valve automation trainer of claim valve cylinder proximity sensor lab experiment output machine cycle trainer program power input output module processor power supply input module output device output jack solenoid input jack automation trainer simulate light student programmable automation controller present invention invention plurality of output jack simulated machine cycle plurality of input jack plc programmable logic controller input device proximity switche such sensor real world machine associated proximity sensor example output solenoid red banana jack indicator light machine black banana jack plurality used push button automation trainer comprising the invention exemplary embodiment embodiment provided plurality

正解 an automation trainer is useful for students to develop programs then download to a programmable automation controller pac or a programmable logic controller plc . the programs can be used to sequence cylinders that are controlled by valves and sensors . the automation trainer along with lab experiments simulate real world problem solving and programming . the automation trainer simulates real world machines and is easily expandable and flexible . cross reference to related application this application claims the benefit of priority of u unk . provisional application no . filed apr . the contents of which are herein

出力 an automation trainer is useful for students to develop programs then download to a programmable automation controller pac or a programmable logic controller plc . the programs can be used to sequence cylinders that are controlled by valves and sensors . the automation trainer along with lab experiments simulate real world problem solving and programming . the automation trainer simulates real world machines and is easily expandable and flexible . cross reference to related machines and programming . the automation trainer simulates real world machines and is easily expandable and flexible . cross reference to related apr . the contents

[2] 失敗例

入力 system electric power distribution system load incidence matrix logic plc example switche feeder system restoration control logic fault power substation restoration control feeder automation logic switch one soft plc product system configuration soft plc plug in method user isolation control logic system configuration incidence matrix based automation logic third substation automation isolation control computer implemented method of claim such connectivity matrix power restoration restoration logic soft plc software soft plc project load node restoration master controller automation system isolation switch isolation logic soft plc second substation available power restoration substation restoration path feeder network search power restoration path computer

正解 a soft plc technology based computer implemented method includes updating a system configuration incidence matrix for an electric power distribution system based on both a depth first search of a connectivity matrix for the electric power distribution system and information about the electric power distribution system wherein the information includes at least status information about one or more switches of the electric power distribution system . the method further includes detecting a fault in the system based on the incidence matrix . the method further includes generating isolation control logic based on the incidence matrix and isolating the fault

出力 systems enables and method for system to system and or the power power power power power power system including a power power power power or a power power power power power power power power system and system and or the power system and or the power and the power and the power power and the power and the power and the power and the power system and the power and the power power and the power and the power and the power and the power and the power and the power and the power system and or the power power power and

6 おわりに

[1] 普通の seq2seq を実装したので提案手法バージョンのモデルを実装する

[2] 入力パラメータとして keywords を入れた方がいいか検討

参考文献

[1] Ministry of Public Management, ”Boost Open Data Strategy”, http://www.soumu.go.jp /menu_seisaku/ictseisaku/ictriyou/opendata/, Accessed: May 5, 2018.

[2] NTT Communications, ”Report of Deploying Information Infrastructure of Hazzard Data”, ”No. 1.0, 2013.

[3] Ilya Sutskever Oriol Vinyals Quoc V. Le, ”Sequence to Sequence Learning with Neural Networks”, *NIPS*, 2014

[4] 斎藤 康毅, ”ゼロから作る Deep Learning 2 自然言語処理編”, オライリー・ジャパン, 2018.