



特許情報に関する言語生成モデルを 活用した知的財産創造手法の開発

Development of Intellectual Property Creation Method
Using Language Generation Model on Patent Information

Shigeaki Onoda

Graduate School of Information Engineering, Toyama Prefectural University
t855005@st.pu-toyama.ac.jp

Friday., November 8, 2018,
Toyama Prefectural Univ.

- 1.
- 1. はじめに
- 2. 特許生成 - モデル
- 3. 知財創造のための変換モデル
- 4. 知財創出手法の提案
- 5. 数値実験ならびに考察
- 6. おわりに
- 質問表
- Appendix



1.1. 本研究の背景

背景

ICT 分野の発達により、民間団体や政府機関のデータをデジタル化することの重要性が増している¹。様々な分野で科学技術による効率化が図られているのに対して、特許分野も例外ではない。日本の特許庁に提出された特許や実用新案等を誰でも容易に検索・照会可能なサービスの一つとして特許情報プラットフォーム²がある。

特許のオープンデータについて

- 1** 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
- 2** 人工知能の第三次ブームにより特許への応用事例もある
- 3** しかし 現状では特許のような非構造データにおける絶対的な処理手法は存在しない

¹http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/

²<https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage>



1.2. 本研究の目的

目的

本研究では複雑な特許のデータを利活用し意思決定の支援となるべく特許データを用いた知財提案システムを作成する。

先行研究

- 1 てがかり表現による技術課題の抽出 (酒井ら 2009)
- 2 テキストマイニングと DEA を用いた新しいビジネス領域の特定 (Seol ら 2011)
- 3 ランダムフォレストを用いて特許文書から技術の適用領域を抽出する手法 (津村ら 2017)
- 4 深層学習を用いたパテントマップ自動生成 (太田 2017)

本研究の枠組み

特許の文書と引用数等のパラメータを複合的に考慮している研究は少ない...

1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
Appendix



2.1. 知財創造とは

1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix





2.1. 本研究が目指すシステム



Figure: 1: システムの概念図

1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix



2.2. 言語生成モデル

知財創造の要件定義

ユーザが求めている知財案を提示し、ユーザの意思決定・発明を支援するシステムに必要なものとして、

- 1 ユーザが求めている分野・分類を反映できるインタフェース
- 2 ユーザが理解できるように出力した新知財案

検討する手法

以上を実現するため、文書から文書を生成するモデルエンコーダー・デコーダーモデルを採用する。

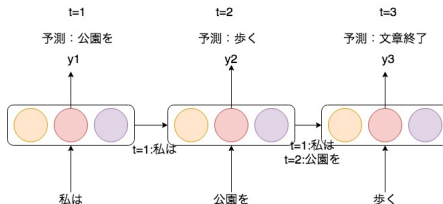


Figure 2: 言語生成モデルの概念図



2.3. 生成モデルと LSTM

生成モデル

言語モデルは過去生成された単語から次に生成される単語の生成確率を確率モデル化したもの。各単語の同時確率が文自体の生成確率を表す。

$$P(Y) = \prod_{t=1}^T P(w_t | w_1^{t-1})$$

Y : 生成文, w_t : t 番目の単語, w_1^{t-1} : $t-1$ 番目までの単語群

言語生成への応用

確率モデルの導出を NN の一種である LSTM³ を用いたものを使うのが主流。

³Long Short Term Memory



3.1. 系列変換モデル

seq2seq

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。シンプルなモデルで Encoder と Decoder からなる LSTM である。Encoder で一旦入力系列の情報を一つの隠れ層の値として吐き出し、それを Decoder の入力にするだけである。

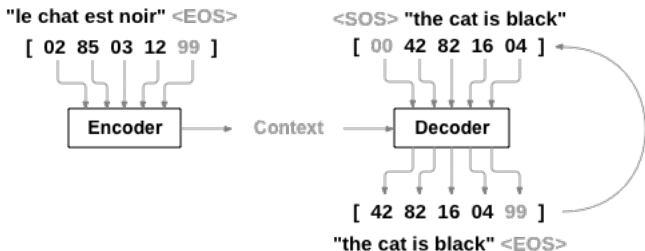


Figure: 3: フランス語から英語翻訳ネット



3.2. seq2seq の詳細

seq2seq の定式化

$$\log p(y|x) = \sum_{j=1}^m \log p(y_j | y_{<j}, x, \mathbf{s})$$

$x_1, \dots, x_n$: 入力系列, $y_1, \dots, y_n$: 出力系列, $\mathbf{s}$: 隠れ状態

seq2seq のふるまい

入力系列を出力系列に写像する条件付き確率をモデル化している。
具体的には, Encoder で一旦入力系列の情報を一つの隠れ層の値として出力し, それを Decoder の入力にするというモデル。

利点: 入力された時系列データを別の時系列データに変換することが可能である点である。補足: 任意の長さの入力を固定長のベクトルに変換する。

欠点: 入力文書の長さ問わず固定長ベクトルに変換するので, 長い文書の場合情報損失がある。



3.3. エンコーダー・デコーダーモデルの応用

画像分野ではエンコーダーに CNN を用いて画像からキャプションを生成するモデルが考案されている。このようにモデルを抽象化することで、系列データ以外の画像にも応用できている。

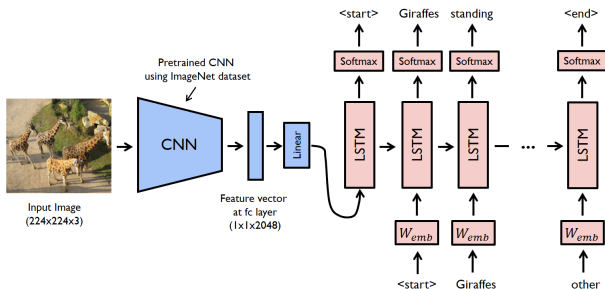


Figure: 4: 神経回路網による自動画像キャプション生成



4.1 特許のデータ収集

特許情報の収集・分析基盤

本研究では収集の対象を **Patents Google⁴** として独自にクローラーを開発した.

- 1 非同期リクエストにより高速化
- 2 NoSQL に保存することでスケーラビリティを確保

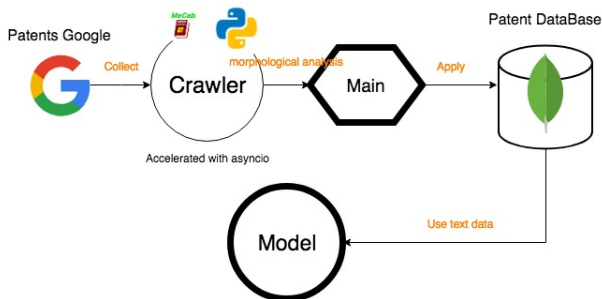


Figure: 5: 収集・分析プラットフォーム



4.2. 特許の価値の抽出

特許の価値の定義

特許の価値として挙げられるものは以下である (後藤ら 2006)

- 1 引用件数
- 2 被引用件数
- 3 論文からの引用件数
- 4 論文からの非引用件数
- 5 その分野の発明者数

特許情報の利用

本研究はこの内オープンデータから集められる引用件数, 被引用件数を対象としてモデルの説明変数として適用した.

本研究での価値

後藤らは特許ごとの価値のみに焦点を当てたが本研究は各単語ごとの価値も算出できるようにした.



4.3 提案手法の概要

コンセプト

エンコーダー部分を改良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。

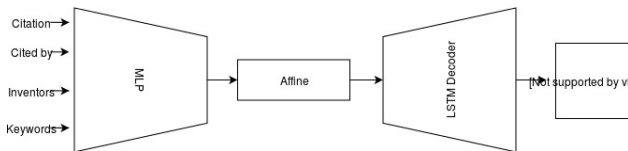


Figure: 6: 新しい提案モデル

提案モデルの仕様

- 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
- 特許の複雑なパラメータ情報を特徴マップとして出力
- 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を行う



5.1. 特許の価値を含んだ単語の重み付け

電灯分野 H05B の特許文書から価値と単語の出現頻度という入力データを扱うデータ包絡分析 (DEA) を用いて特許価値を含んだ重み付け算出した。

Table: 1. 重みの高い上位 10 単語の抜粋

Word	Weight	Frequency
アントラセン (anthracene)	5.855055	3
下方 (belowdown)	2.343041	26
クラック (crack)	1.785240	10
さ (unknown)	1.721771	117
光 (light)	1.603364	948
アルミニウム (aluminium)	1.405620	33
システム (system)	1.345943	70
エネルギー (energy)	1.070877	31
アノード (anode)	0.999145	29
お呼び (invitation)	0.920917	218

従来の単語頻度とは異なり、電灯分野特異な専門用語に高い重みづけがなされている。この単語重み付けを用いることで**価値を考慮したモデル**が作成可能だと考えられる。



5.2. LSTM を用いた言語生成

実験設定

コーパスから単語の出現を学習したモデルを用いて文を生成する。

Table: 2. モデルのパラメータ設定

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

使用データ

automation と quantum computer の 2 分野 1651 記事

価値の反映法

前述の価値を含んだ重み付け手法, もしくは分散分析を用いて数値化

- 1.
 1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix



5.3 実験結果

実験結果上述のデータを用いて特許のキーワードから特許の文の生成を行った. 表3のような結果になった. 学習データはまだ少ないがエポックを多めにしたため約 50 時間かかっている

Table: 3. 学習の結果

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

- 1.
 1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix



5.4 学習結果の一部抜粋

以下に学習結果の抜粋を載せる．対象の文書は長文のため別ページにて記載

<https://hackmd.io/VujQDs47TRqjGiAARIFD6g>

考察

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ます．ただし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.
- 1. はじめに
- 2. 特許生成 - モデル
- 3. 知財創造のための変換モデル
- 4. 知財創出手法の提案
- 5. 数値実験ならびに考察
- 6. おわりに
- 質問表
- Appendix



5.4

参考：日本語訳

オートメーショントレーナーは、学生がプログラムを開発し、プログラマブルオートメーションコントローラー pac またはプログラマブルロジックコントローラー plc にダウンロードするのに役立ちます。このプログラムは、バルブとセンサーによって制御されるシリンドラのシーケンスに使用できます。自動化トレーナーとラボの実験は、実世界の問題解決とプログラミングをシミュレートします。自動化トレーナーは実際のマシンをシミュレートし、簡単に拡張および柔軟にできます。関連するマシンおよびプログラミングへの相互参照。自動化トレーナーは実際のマシンをシミュレートし、簡単に拡張および柔軟にできます。関連する apr への相互参照。

- 1.
1. はじめに
2. 特許生成 - モデル
3. 知財創造のための変換モデル
4. 知財創出手法の提案
5. 数値実験ならびに考察
6. おわりに

質問表

Appendix



5.5 評価指標

論文や文献を調査した結果、本実験のような生成タスクではアンケートによる定性的評価を用い
定量的評価では一般的に ROUGE⁵ が用いられる。

$$ROUGE_N = \frac{\sum_{S \in references} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in references} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

* S : 生成文の n -gram 集合 * $Count_{match}$ 予測と人手文中の n -gram における最大一致単語数

概要

これは N-gram 単位の一一致を取る指標で Recall と Precision がある

- Recall: 人手の文中の単語を、どれだけ当てられたか
- Precision: 生成した単語中どれだけ人手の文の単語が存在するか

⁵<https://www.aclweb.org/anthology/W04-1013.pdf>



6. おわりに

今後の展望・課題

- 長文における end-to-end モデルの手法を見つける
- 技術分野の記事をもっと集める
- どのように多入力エンコーダーから一つの特徴量に写像するかを考えること
- コンジョイント分析との融合でどのように属性を提示するか検討



ご清聴ありがとうございました。



1.

1. はじめに

2. 特許生成 - モデル

3. 知財創造のための
変換モデル

4. 知財創出手法の
提案

5. 数値実験ならび
に考察

6. おわりに

質問表

Appendix

質問例をまとめました. ご参考ください.

- なぜ生成モデルを使うのか?
- なぜルールベースでやらないのか?
- 特許のデータはどのようなものか?
- LSTM とは?
- 使っている形態素解析辞書は?



言語生成モデル

1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表

Appendix

言語生成のプロセス

- 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

生成モデル

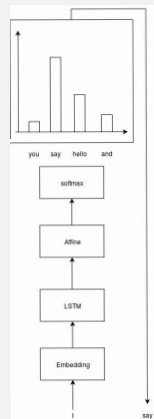


Figure: EX1: 生成モデル



言語生成モデルの具体的な説明

言語生成モデルで使われるネットワークはエルマン再帰型ネットワークの改良型の LSTM が用いられる

エルマン型再帰型ネットワーク

文書や株価の情報等の時系列なデータを処理する際に有効なニューラルネット (NN) の派生系下の図のように前の時間の重みを受け取る再帰的構造を持つネットワーク

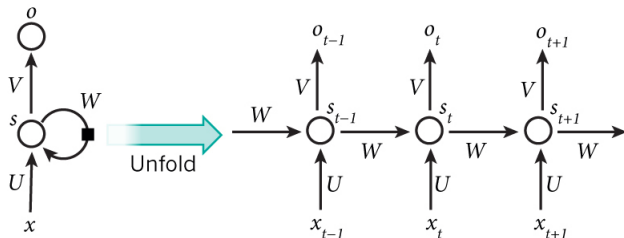


Figure: EX2: RNN



RNN の定式化

- 1.
 1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix

定式化

$$h_t = \tanh(h_{(t-1)}W_h + x_tW_x + b)$$

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

RNN の弱点

RNN を用いて時系列の依存関係を学習できた. しかし, かなり前の情報も考慮していたため長期の依存関係を学習する際は勾配爆発や勾配消失が起こる可能性があった



LSTM

ゲートと記憶セルという仕組みを導入して RNN の弱点を改善したのが LSTM(Long Short Term Memory) である

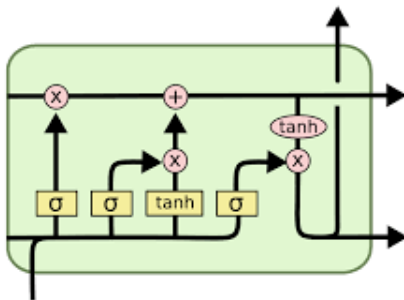


Figure: EX3: LSTM ゲートの概念図

画像引用: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>



1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表

Appendix

定式化

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

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



- 1.
 1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix

$$\begin{aligned} \max : \quad & y_k = \sum_{n=1}^N v_{kn} y_{kn} \\ \text{subject to : } & \sum_{m=1}^M u_{km} x_{sm} - \sum_{n=1}^N v_{kn} y_{sn} \geq 0 \\ & \sum_{m=1}^M u_{km} x_{km} = 1 \quad (s = 1, 2, 3, \dots, K) \\ & u_{km} \geq 0 \quad (m = 1, 2, 3, \dots, M) \\ & v_{kn} \geq 0 \quad (n = 1, 2, 3, \dots, N) \end{aligned}$$



コンジョイント分析との融合

- 1.
1. はじめに
2. 特許生成 - モデル
3. 知財創造のための変換モデル
4. 知財創出手法の提案
5. 数値実験ならびに考察
6. おわりに

質問表

Appendix

改善案

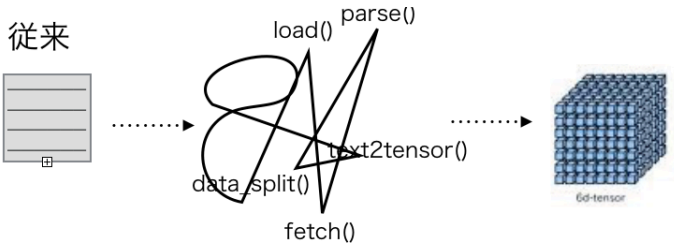
前述の NN では入力に引用数・キーワード等の情報を入れてあるが、これでは意思決定者が利用するインタフェースとして不透明である。そこで、コンジョイント分析を用いて意思決定者の好みのパラメータを同定して NN インタフェースに受け渡す仕組みを作る必要がある

コンジョイント分析とは

主にマーケティングの分野で使われる分析手法、商品の値段、スペック、大きさ等のパラメータ (属性) の最適な組み合わせを提示する手法である。



- 1.
 1. はじめに
 2. 特許生成 - モデル
 3. 知財創造のための変換モデル
 4. 知財創出手法の提案
 5. 数値実験ならびに考察
 6. おわりに
- 質問表
- Appendix



シンプルかつ柔軟なデータパイプラインを構築

Figure: 1: データローダーの概要