修士論文

1-発表番号 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

富山県立大学　電子・情報工学科
1855005　小野田 成晃
指導教員　奥原 浩之 教授
令和2年2月19日

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号
先行作業	i
後続作業	j
プロジェクトの総作業数	n
従事者グループ	k
依頼候補の従事者グループ数	w
作業の所要時間	t_{ij}
作業を従事者グループに依頼したときの費用 (ファジィ・ランダム変数)	$\tilde{\tilde{c}}_{ik}$
クリティカルパスを選択する 0-1 変数	x_{ij}
依頼する従事者グループを選択する 0-1 変数	y_{ik}
ファジィ・ランダム変数のメンバシップ関数	$\mu_{\tilde{\tilde{c}}_{ik}}$
中心値 (平均値)	$\bar{d}_{ik}$
左右の広がりパラメータ	β_{ik}, γ_{ik}
ファイジィ目標	$\tilde{G}$
ファジィ目標の最良値	g^0
ファジィ目標の最悪値	g^1
ファイジィ目標のメンバシップ関数	$\mu_{\tilde{G}}$
ファジィ目標は満たされる可能性の度合い (可能性測度)	$\Pi_{\tilde{\tilde{c}}_{ik}}$
満足基準値	h
擬逆関数	$L^*(h)$, $\mu^*_{\tilde{G}}(h)$
確率変数の分布関数	F

第1章

はじめに

§ 1.1　本研究の背景

ICT 分野の発達により, 民間団体や政府機関のデータを 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.

特許情報には過去の発明の情報をアーカイブしたいわば発明の保管庫的なデータであり, それを利活用することで経営戦略・技術的発展等広く社会に役立てることができる. このようなオープンデータやプラットフォームを用いて多角的な特許分析を行うことが特許分野ひいてはデータマイニングにおける重要な貢献といえるだろう.

しかし, 現状の特許プラットフォーム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. そのため現状の特許分野では日英翻訳タスク等の主な特許タスクでは国立情報学研究所の実験用コレクション NTCIR[2]が用意されているものの一部な高度な分析においては各々が独自にデータを取得する必要がある.

特許分野における情報技術の応用例としては以下の 3 つに大別される [3].

[1] https://www.j-platpat.inpit.go.jp/web/all/top/BTmTopPage
[2] http://research.nii.ac.jp/ntcir/data/data-ja.html

特許情報処理システムの分類

- 業務推進系システム
- 管理系システム
- 分析評価システム

これらの技術を用いることで特許出願・分析における業務の効率化を行うことができる.

特許の研究については多言語間特許の翻訳や統計学を用いた特許分析が活発に行われている. ここで特許データはテキストと引用数等の表形式が混同する非構造化データである. このような複数のモダリティを考慮した研究は上述した研究に比べ少量である.

また, 管理系システムに分類される発想支援システムは KJ 法や TRIZ 等の発想支援論を元にシステムを構築するものが大半であり, 具体的な特許データに基づいた数理的なシステムは少ない.

このことから，特許の複数パラメータを複合的に考慮しつつ, 意思決定支援となる発想支援システムの新たな枠組みが必要となる.

§ 1.2　本研究の目的

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活用するための発明支援システムを提案する.

システムを開発するために, 特許データを効率的に収集・保存する枠組みを考案し, そのデータを元に特許の価値やキーワードの特徴を捉えたモデルを作成する. 最終的にそのモデルを利用してユーザが使えるようなインタフェースを含んだ実用的な発想支援システムの開発を試みる.

具体的には各特許の情報及びその関係性を n 次元空間にマッピングすることができればそこからコサイン類似度を求めて類似特許を求めたり, 足し算等のベクトルができる演算可能になる. これにより特許を俯瞰的に分析しつつも新しい疑似特許生成してアイデアの補助となるキーワードを提示することが可能になる.

システムの実証としては定量的評価と定性的評価の 2 つの側面から行う. 定量評価としてはモデルが特許特性をどれほど反映できているかを検証し, 定性評価では実際に人にシステムを利用してもらいフィードバックを得ることでシステムの有用性を確認する. これにより, 数理的な優位性を持ちつつも人が利用しやすいシステムになっているかを検証する.

§ 1.3　本論文の概要

本論文は次のように構成される.

第 1 章 本研究の概要と目的について説明した.

第 2 章 従来の特許情報処理と発想支援システムについて述べ，それらの有用性と課題について述べる．

第 3 章 システムで利用する手法とその前提となる重要な技術や概念をここで紹介する．

第 4 章 提案手法とそれを用いたシステムのアーキテクチャについて説明する．

第 5 章 システムに対する実験結果と考察を述べる．

第 6 章 まとめと今後の課題を述べる．

第2章

特許情報処理とアイデア発想支援

§ 2.1　特許情報処理

特許情報処理とは

特許とは知的財産の一種であり発明の保護を目的として, ある発明に対して独占的・排他的にジッチするために付与された権利であり, 特許権とも呼ばれる. 日本では約 34 万件の特許が出願され, 多様な分野の発明が蓄積されている [3].
また, 特許の構成は自然言語のテキストと補足的な情報と図表で構成される. そこで特許分野に対して自然言語処理の技術を利用することにより, 産業上の価値を産出するのに役立てることができる.
そして, 特許情報処理の技術を工学的に評価し検証を行うことにより学術的・産業的に意義のある成果をもたらすであろう.

本研究における特許の範囲

特許とは前述の知的財産としての意味もあれば, それ自体を申請する行政処理も含む場合もある. 特許が含有する特許情報としては, 主に「特許請求の範囲」, 「明細書」, 「要約」からなる. 本研究は専ら特許情報を工学的応用に主観を置くため, 上記の特許情報を対象とする.

特許の一例

本研究で利用する特許の実例を図 2.1[1]に付す. このように特許はタイトルと要約である **Abstract**, IPC という特許分類を示す **Classification**, 本文を示す **Description**, 請求項を表す **Claims**, そして特許自体の ID と出願日などの情報を含んだ部分からな

[1] https://patents.google.com/patent/WO2006086021A2/en?oq=robots

る．また図には示されていないが特許の引用の情報も検索プラットフォームによっては示されている．

図 2.1: 特許文書の一例

§ 2.2　特許情報処理システム

第 1 章で示したように特許は「業務推進系システム」,「管理系システム」,「分析評価系システム」と大別される．これらの概要と事例について説明する．

業務推進系システム

このシステムは主に特許の戦略立案や発明着想, 調査ひいては特許無効化のために利用される．また, ここには以下の 6 種類のシステムが含まれる．

- 「特許検索」
  類似特許がないか調べるための
- 「パテントマップ作成」
  なんらかの形で収集された特許からパターンや年代を分析して目的に応じた形で出力する
- 「発想支援」
  発想支援理論を用いて知財考案時におけるアイデアを整理・刺激する
- 「明細書作成支援」
  特許出願における明細書等の文書の作成を半自動で行い効率化を目指す

- 「特許翻訳」
  外国語特許に対して検索・分析する際に利用される
- 「出願支援」
  特許のオンライン申請を可能とするインターネット出願ソフト[2]

管理系システム

特許実務において一つの特許に対して様々な文書が付与される. そしてそれらは随時更新される. そのような様々な文書を長期に渡って保守管理を行うためのシステムがここに含まれる. この分野では主にデータベースや保守方法等が議論されるため本研究では対象としない.

分析評価系システム

この分野では主に特許の情報を統計学や機械学習を用いて分析する. 特許作成の支援を直接行うのではなく, 学術・産業上の目的のために利用される. 主に以下の 3 つに大別される.

- 「特許分析」
  複数間特許を横断的に分析することで, 企業の勢力図や現在の技術の流行等をみつける手がかりを得る事ができる
- 「特許明細書分析」
  単一の明細書に対して構造解析を行い, 読解支援や品質評価に用いられる
- 「特許価値評価」
  ある特許が持つ価値の算定を目的とする. 経済的評価に偏重しており技術的価値の評価の研究は途上である

特許分野の研究事例特許分野の研究として上記でいう「情報検索」,「特許翻訳」,「パテントマップ作成」, 「特許分析」が活発に行われている. 例えば, Mase らの研究で特許文書の温度等の数値表現に大きな重みを与えて TF-IDF を構築して特許検索を行った [4].

そして人工知能の第三次ブームにより人工知能の多方面の応用研究が盛んになっており, 特許も例外ではない. 実用化された特許調査ツールには, トーマツデトロイト社が開発した PLSA をもとにした特許調査ツール Doloitte Analytics がある. また日本においても FRONTEO 社と TTDC: Toyota Techno Developement 社が共同で開発した KIBIT が市販されている [5].

パテントマップを数理的に作成する手法としては, Young らの対象となる技術分野の単語から特許の文書を k-means を用いてクラスタリングするものがある. これはクラスタリング後の結果から semantic network を構築し可視化を行う手法である [7]. また調査対象の特許分野の少数の特許を入力とし, それに対してランダムフォレストを用いることにより特許の集合から目的に合致するサンプルを選択する. そしてそのサンプルの文字データに対して自然言語処理により課題表現を抽出してパテントマップを作成するというものもある [6].

[2]https://www.pcinfo.jpo.go.jp/site/3_inet/index.html

また, 酒井らのように特許の文書の傾向を分析して自然言語処理的なアプローチでの技術課題の抽出手法も提案されている [8]．例えば「ができる.」等の手がかり表現を含む文節に課題表現があるということが判明している．

しかし，これらの研究は特許の文書以外の情報，例えば特許分類，引用件数や発明者・企業等の情報を考慮していない. そこで, 津村らは特許中の出現単語と特許の分類の両方を結合したデータを作成するために，ランダムフォレストによる特許分類の学習を行った．その出力結果に対して MDS を用いたマッピングを行い，特許の分類情報を含んだ特許間の類似度空間を構築した [9].

また, 特許分野に関して経営工学の手法を用いた Hyonju らの研究がある [10]．Hyonju らは企業と産業領域を対象としてそれらの類似性を発見することを目標とした．そしてテキストマイニングと経営工学の手法をあわせてマルチモーダルな単語空間の作成して単語ごとの特許においての技術的価値を算出した Onoda らの研究がある [11].

これらの特許情報処理システムの内「発想支援」の研究では KJ 法等のアイデアを整理するための発想支援論に基づくものが大半であり大規模な特許データをもとに数理的・統計的な知見を用いて作成されたシステムは少ない.

つまり, この分野の研究を行うことで, 産業利用において有用な手法・システムを提案できる可能性が大いにあると思われる. そこで, 次節では特許以外の分野も含めた発想支援事例・研究について詳細に紹介していく.

§ 2.3　発想支援システム

発想支援とは発想支援とは人々が何らかの意思決定やアイデアを出す際にその糸口となるように人間の思考を補助する方法論である．有名なものとしては文化人類学者の川喜が提案した KJ 法がある [12].

KJ 法

KJ 法は図 2.2[a]のように思いついた事象をカードに記し, それらを複数枚並べカテゴリごとにグループ化してそれぞれのグループそしてそのカード間の関係を矢印や線を使い可視化する．これにより雑多なアイデアを効率的に整理することができる．また KJ 法を支援するソフトウェアも作成されている [13].

[a] http://www.ritsumei.ac.jp/se/rv/yamada_s/tea/env/index.html

しかし KJ 法等発想支援論を用いる方法では知識データを利用しておらず, 各アイデア自体は人間が考える必要がある. そこで伊藤らや西原らは電子データを用いたアイデア処理を提案している [14][15].

伊藤らはユーザがブレインストーミングした単語に対してクラスタ分析を行い, ユーザのアイデアが含まれるクラスタと共起度が低いクラスタの単語を提示することで新たなアイデアを刺激するようにしている.

対して西原らの研究では人間が発想した 2 つの商品を組み合わせたアイデアについての


若者がにぎわう施設
観光事業の活性化
(情報発信)
車も人も安全に通れる道路
景観の保全
学校を利用した活動
商業施設の充実
日常生活
伝統文化の商品化
地域に密着した活動
河川とまちの融合

図 2.2: KJ 法の利用例

希少性・新規性を評価する仕組み作った．この研究では膨大な商品アイデアに対する評価をランキング付して提示することで商品創案の支援を行っている

　このように既存のシステムの有効性は示せてるものの, 依然として人間が発想の種となるブレインストーミングが必須となっている．そこで著者は自然言語処理の一分野である言語生成を用いることでより高度なシステムが構築できると考えた．次章では自然言語処理における言語生成について説明する．

第3章

言語生成

§ 3.1　自然言語処理と言語生成

まず, 言語生成とその元となった自然言語処理の基本事項について説明する. **自然言語処理 (natural language processing, NLP)** とは我々が日常用いている自然言語に対して, ある種のコンピュータ処理とそれに関する学問分野, 研究分野をそう呼ぶ [16].

言語というもの大きく以下の特徴を持っている.

言語の特徴

- 物事に対してどう解釈するかは恣意的である
- 論理的に説明できるものだけでなく社会の慣習を反映したものも多い
- 語彙や用法は時代とともに変化し, 専門分野ごとに異なる意味合いを持つ
- 言語の意味内容はネットワーク構造をもつ
- 表現は多対多である. つまり多義性・曖昧性をもつ.

このように言語とは様々な記号や要素が複雑に絡み合い形成されるものであるため, それを機械で解釈することは AI 完全 (AI-complete) であり AI で解くことが困難な問題の一つである [17].

また, NLP には複数の小分野があり, 国や業界により解釈が異なるが日本の場合, 図 3.1[1] のように分けられる. それぞれの分野において異なる手法が取られる場合もあるが, BERT(Pre-training of Deep Bidirectional Transformers for Language Understanding) のように複数のタスクを実行できる汎用的なモデルも考案されている [18].

[1] https://www.zenknow.tokyo/entry/2018/09/07/200849

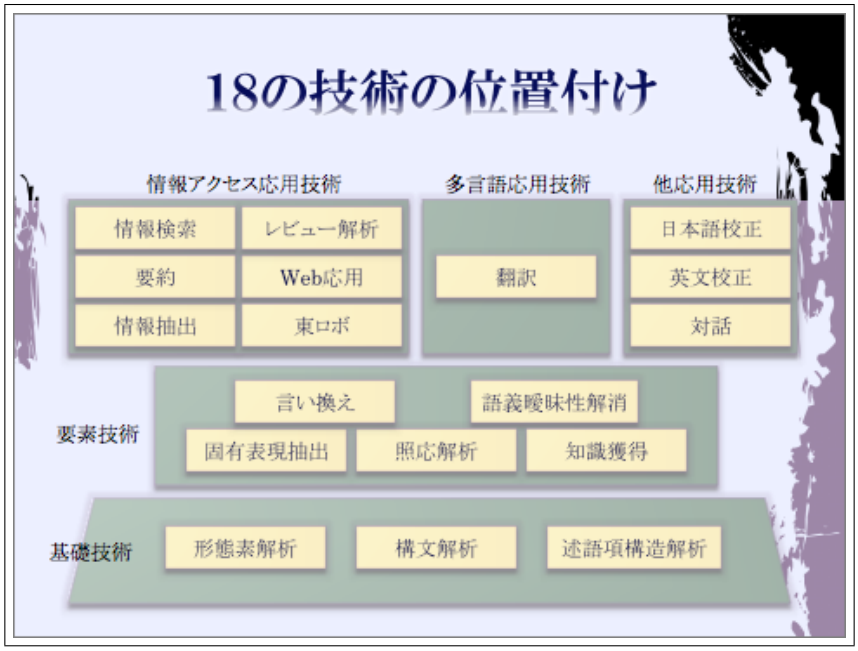


図 3.1: NLP における分野一覧

これらの分野のうち言語生成に特に関わっている技術としては, 知識データを利用する情報アクセス応用技術, 言語生成の関連分野である対話になる. 対話システム自体, 入力から受け取って解釈した後, 知識データと文や言葉の確からしさを表現するある種の確率モデルである言語モデルを横断的に処理して出力を生成する. この一連のフレームワーク自体は機械翻訳や発想支援等の関連分野でも同一である.

そこで次節では対話システムや機械翻訳における基礎技術・理論を順に説明していくことで, 本研究で重要な位置を占める言語生成の手法について理解を深めていくことにする.

§ 3.2　言語生成のための理論・技術

対話システムの初期の代表的なものでは ELIZA と SHRDLU がある [16]. これらは if-then を用いたルールベースのシステムである. このシステムは図 3.2[2] のように CUI ベースで利用できるものである. 例えば always を含む入力に対しては「Show me some specific examples」と出力するといったルール群を用意することで限定的ながらも会話が可能なシステムである.

しかし, このようなルールベースのシステムでは実用的ではない, なぜなら入力に対する適切なルールを人間が手動でセットする必要があり, さらに会話において適切な答えは文脈やコンテキストによって大きく変わるからである.

2019 年論文執筆時ではそのようなルールベースではなく文としての確からしさを大規模なコーパスのパターンを統計学・機械学習を用いて学習する技術を言語モデルが使われている.

[2] https://en.wikipedia.org/wiki/ELIZA#/media/File:ELIZA_conversation.jpg

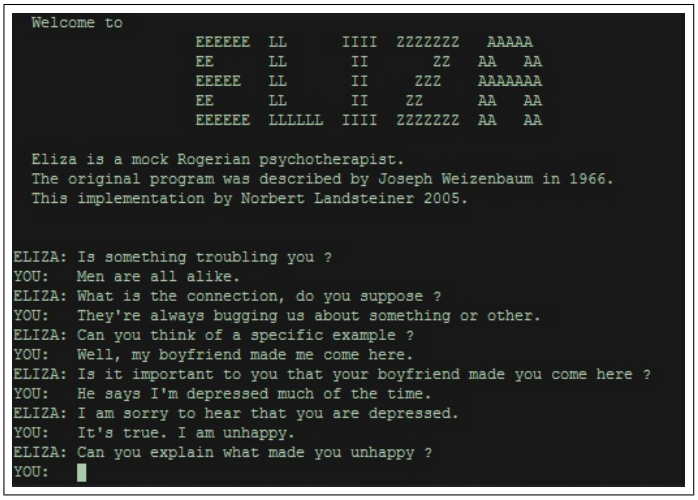


図 3.2: ELIZA の利用例

3.2.1 言語モデル

言語モデルは文や文書の生成確率モデルであり, $w_1w_2,\ldots,w_n$ という言語表現に対して, 確率 $P(w_1w_2,\ldots,w_n)$ を考える場合以下のようにモデル化する [19].

$$P(w_1w_2,\ldots,w_n) = P(w_1,w_2)P(w_3|w_1,w_2)P(w_4|w_2,w_3)\ldots P(w_n|w_{n-2},w_{n-1}) \tag{3.1}$$

この内のパラメータ $P(w_i|w_{i-2},w_{i-1})$ をいかに精度を高く求めるかを研究が行われている. この言語モデルの学習方法についてはマルコフ連鎖によるもの, ニューラルネットを用いたものなど様々なものが存在する.

隠れマルコフモデル

古典的な言語モデルの利用例としては隠れマルコフモデル (hidden Markov model, HMM) によるものがある. 例として対話システムをモデル化する場合を考える.

ある特定テキストデータを単語ごとの系列と考えると以下のようにモデル化できる. HMM とは入力系列 $\boldsymbol{x}$ とした場合, 出力系列 $\boldsymbol{y}$ をとすると, HMM は各状態の直前の状態のみに依存する. x_i は y_i のみに依存して, y_i は y_{i-1} のみに依存すると仮定すると $\boldsymbol{x},\boldsymbol{y}$ の同時確率は以下のように表せれる [19].

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{y}) &= P(x_k,y_k|x_{k-1},yk-1)P(x_{k-1},y_{k-1}|x_{k-2},y_{k-2})\ldots \\ &\quad (x_2,y_2|x_1,y_1)P(x_1,y_1|x_0,y_0) &(3.2)\\ &= \prod_i P(x_i,y_i|x_{i-1},y_{i-1}) &(3.3)\\ &= \prod_i P(x_i|y_i)P(y_i|y_{i-1}) &(3.4)\end{aligned}$$

ここでダミー要素 (x_0,y_0) を用いて式を変形した.

この式に対して最尤推定を行うことで入力系列と出力系列の対となるデータセットから各入力系列ごとに尤もらしい生成単語の確率分布を出力する. 確率分布から実際の出力単語の選択方法としては貪欲法やグリッドサーチ等のアルゴリズムが用いられる.

このように HMM を用いることで系列データを扱うことが可能になった反面, 課題もある. 言語データは直前の単語だけでなくいくつか前の単語や前の文の内容も影響しているため, それをモデル化する際は当然考慮すべきである. しかし HMM の場合一つまえの状態のみしか考慮していない.

そして HMM は機械翻訳や対話システム等多くの NLP の諸分野でより表現力が高く一般的なリカレントニューラルネット (Recurrent Neural Network, RNN) にとって変わられた. RNN については次の節で紹介する.

RNN の概要

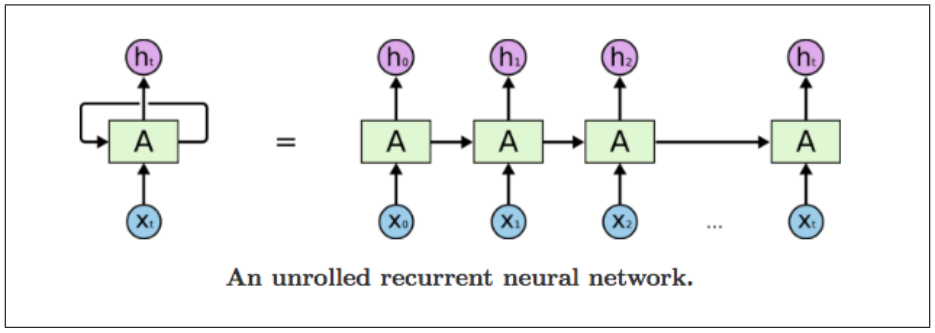


図 3.3: RNN の構造

RNN とはエルマンによって考案されたニューラルネットの一種である. 通常のニューラルネットと違い図 3.3[3]のような再帰性を持つ. ここで A は一つのニューロン, x_t がある系列の入力そして h_t がその時の隠れ状態 (潜在変数) である.

この再帰性により言語のような系列データに対してチョムスキーが唱えた生成文法のような事前知識を仮定しなくとも文法構造に関する知識が創発することが確認されている [20].

これにより幼児が少ないインプットから言語獲得可能な「刺激の貧困」に対する生成文法とは異なる説明といえるだろう.HMM やそれを一般化した状態空間モデルとの異なる利点は以下となる [21].

[3]https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e

RNN の利点

- 状態（潜在変数）を導入することで時系列を考慮した予測ができる
- 状態空間モデルより一般的
- 長期依存性を持つ
- ドメイン知識不要
- ニューラルネットのため誤差逆伝播が可能

このような利点から機械翻訳から株価の時系列予測まで広く使われている [22][23]. しかし, 万能なモデルではなく以下の問題点がある.

RNN の問題点

- 複雑なモデルのためロバスト性が担保しずらい
- オブザベーション数（データ量）が少ないとモデルの複雑性に対して学習不足になる
- 計算コストが高い

そのため RNN を組み込む際は取り組む課題のデータ量や許容できる計算コストと相談しながら適用することになる. また, RNN は図 3.3 のようにシンプルな構造な反面, 長期の依存関係を学習することが難しく勾配消失や勾配爆発を引き起こすことがある [24]. そこで考案されたのが RNN の拡張モデルである長・短期記憶 (Long Term Short Memory) である.

LSTM

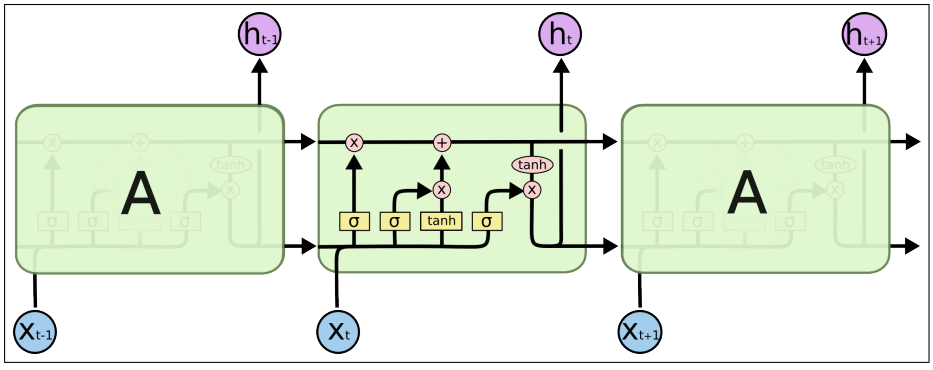


図 3.4: LSTM の構造

LSTM は RNN に対してどこまで過去の記憶を保持するか調整するゲートを設けたモデ

ルである. 図 3.4[4]のように諸演算を行うことによるゲートを数理的に表現している. 定式化は以下のようになる.

$$
\begin{aligned}
\boldsymbol{f} &= \sigma(\boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{x}}^{\boldsymbol{f}} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{\boldsymbol{h}}^{\boldsymbol{f}} + \boldsymbol{b}^{\boldsymbol{f}}) \\
\boldsymbol{g} &= \tanh(\boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{x}}^{\boldsymbol{g}} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{\boldsymbol{h}}^{\boldsymbol{g}} + \boldsymbol{b}^{\boldsymbol{g}}) \\
\boldsymbol{i} &= \sigma(\boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{x}}^{\boldsymbol{i}} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{\boldsymbol{h}}^{\boldsymbol{i}} + \boldsymbol{b}^{\boldsymbol{i}}) \\
\boldsymbol{o} &= \sigma(\boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{x}}^{\boldsymbol{o}} + \boldsymbol{h}_{t-1} \boldsymbol{W}_{\boldsymbol{h}}^{\boldsymbol{o}} + \boldsymbol{b}^{\boldsymbol{o}}) \\
\boldsymbol{c}_t &= \boldsymbol{f} \odot \boldsymbol{c}_{t-1} + \boldsymbol{g} \odot \boldsymbol{i} \\
\boldsymbol{h}_t &= \boldsymbol{o} \odot \tanh(\boldsymbol{c}_t)
\end{aligned}
\tag{3.5}
$$

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

式 (3.5) の利点としては RNN から記憶セル $\boldsymbol{c}_t$ 以外同じインタフェースで利用でき, かつより長文に対して精度の高い予測ができる点である. また上のアファイン変換分 $(\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{i}, \boldsymbol{o})$ は以下のように変形することでまとめて計算して高速化できる.

$$
\boldsymbol{f}, \boldsymbol{g}, \boldsymbol{i}, \boldsymbol{o} = \boldsymbol{x}_t \boldsymbol{W}_{\boldsymbol{x}} + \boldsymbol{h}_t + \boldsymbol{b} \tag{3.6}
$$

LSTM の諸テクニック

LSTM もこのまま利用してもある程度は学習できるがより能率的に学習するための諸テクニックを紹介する [24]. 学習がうまくいかない場合はややトリッキーにはなるが以下のテクニックを試すと精度が上がる可能性がある.

LSTM(RNN) のテクニック

1. 多層化
   LSTM を 2 層以上にすることで通常のニューラルネットのように表現力が上昇する.
   しかしハイパーパラメータ数も増加してモデルが複雑化する.
2. 正則化の適用
   過学習の防止のため正則化項を加える. LSTM の場合は Dropout を深さ方向に適用する.
3. 双方向 LSTM
   テキストに対して有効なテクニック. 単語系列に対して左から右と右から左両方学習させる.
4. 重み共有
   特定の層で重みを共有することで計算量とモデルの複雑性を同時に減らすことができる.

[4] https://colah.github.io/posts/2015-08-Understanding-LSTMs/

§ 3.3 seq2seq

前節で説明したLSTMを用いて言語生成ような系列データから系列データに対して代表的なモデルとして系列変換モデル (sequence to sequence, seq2seq) がある.

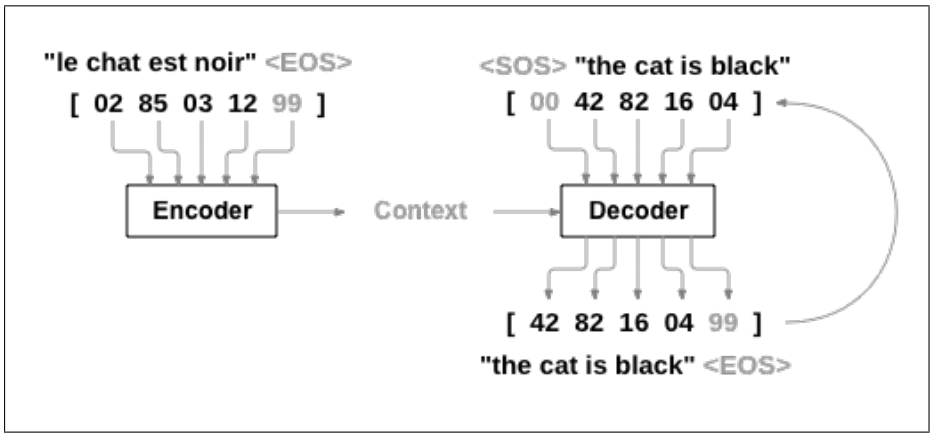


図 3.5: seq2seq の概念図 (英仏翻訳の例)

seq2seq は Sutskever らによって考案されたモデルであり, エンコーダとデコーダという２つの LSTM ネットワークを利用することで英仏翻訳で発表当時最高のパフォーマンスを発揮したモデルである [22].

式は以下のように確率モデルに隠れ状態 s を加えた形で表せる.

$$\log p(y|x) = \sum_{j=1}^{m} \log p(y_i|y_{<j}, x, \boldsymbol{s}) \tag{3.7}$$

$x_1, ..., x_n$: 入力系列, $y_1, ..., y_n$: 出力系列, ￥bms: 隠れ状態

seq2seq の特徴

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.

図 3.5[5]のようにフランス語の文章列をテンソルに圧縮する ENCODER, 圧縮された系列を英語の文章に変換する DECODER からなる. このモデルが以降の研究に大きな影響を与えた. その理由としてはエンコーダ・デコーダという抽象化した枠組みで捉えることを可能にした点である.

これによりエンコーダ・デコーダの中身を様々な形にカスタマイズしてもインタフェースさえ守っていれば正しく系列変換される.

つまり, このモデルは閉鎖・開放原則に従っているといえ, 高い拡張性と変更に強いものになっている. 実際中身を変えた応用研究も散見される [25]-[28].

seq2seq の応用研究として興味深いものとしては Oriol らが考案した自動キャプション生成モデルである [29]. このモデルは図 3.6 の通りエンコーダに画像認識で高い効果を発揮す

[5]https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

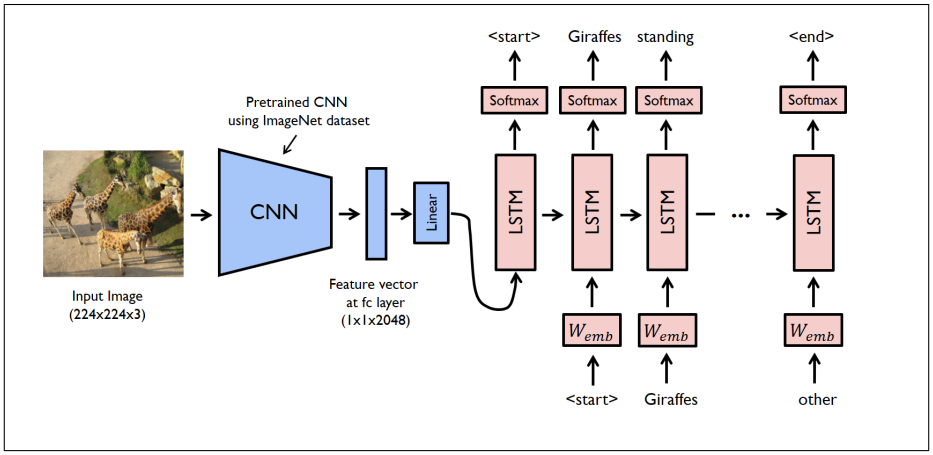


図 3.6: 自動画像キャプション生成ネットワーク

る Convolution Neural Network(CNN) を用いて特徴をベクトル化するそれを LSTM で構成されるデコーダで復元することで画像に対して適切なキャプションが生成される.

§ 3.4 VAE による言語生成

ここまで seq2seq をベースとした言語生成を見てきた. 確かに答えが用意されている翻訳タスクでは実用レベルで使われている. しかし, 本研究が目標とする知財創造においては答えを用意することは容易ではない. そこで, 教師なし学習である自己符号化器 (Auto Encoder, AE) を拡張した VAE(Variational Auto Encoder) を用いて言語生成を行う.

AE

AE は Hinton らが提案したアルゴリズムでデータの次元圧縮にニューラルネットワークを用いたものである. 教師なし学習であるが, 学習対象のデータを入力だけでなく教師データとしても扱うことで学習する.

図 3.7 のように入力に対してその情報を圧縮するニューラルネットワークを用いて一つのサンプルを一つの潜在変数 z に圧縮して, その z をデコーダネットワークの入力として復元する. そして入力サンプルと出力サンプルが近くなるように誤差を修正することで学習する. ここでの誤差関数としては二乗誤差が用いられることが多い.

この学習がうまくいけばいくほどデータの特徴を捉えた圧縮情報が得られる. しかし z の値は人間には解釈が難しい. そこで提案されたのが次に説明する VAE である.

VAE

VAE は AE の z に対して $z \sim N(0,1)$ を仮定したモデルである [31]. このモデルの大きな貢献は 2 点ある. ひとつは, このモデルは変分下限を損失としてパラメータ化することで学習の高速化をもたらした点もうひとつは, 高次元なデータの存在する確率分布を求めることが可能になり, 未学習の新しいサンプルを得ることができることである.

以前でも生成モデルにおいて確率分布に落とし込むこと自体は可能であったが, データ構造への強い仮定やモデルに強い近似が必要であった. VAE によってより容易に解釈性の高い特徴量を獲得できるようになった.

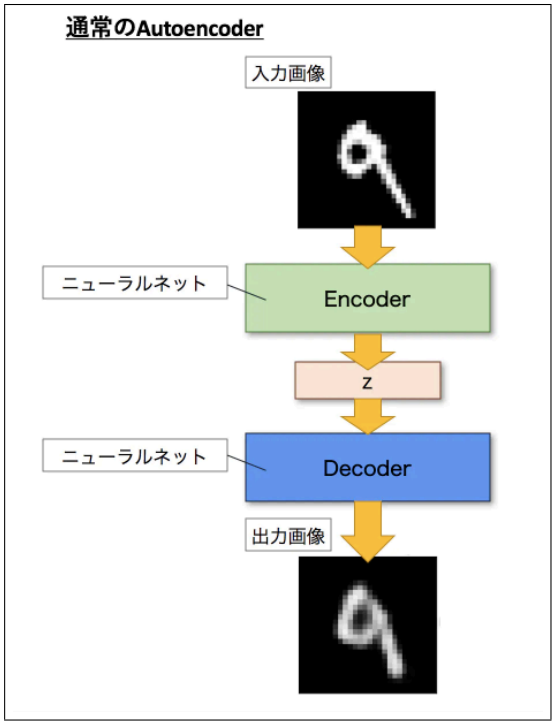


図 3.7: AE の概念図

VAE の定式化

VAE におけるデータの確率分布を求める式は以下のようになる. 通常 $\log p(X)$ を求めることは困難なので変分下限と確率分布の KL ダイバージェンスで近似して求めることで計算コストを下げている.

$$\log p_\theta(\boldsymbol{x}^{(i)}) = D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})|p_\theta(\boldsymbol{z}|\boldsymbol{x}^{(i)})) + L(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{x}^{(i)}) \tag{3.8}$$

ここで確率分布 $p(X)$ を扱いやすいように対数で表す. なお phi, θ は事前に最尤法によって求める.

なお $z \sim N(\mu(X), \sigma(X))$ はそのままだと確率分布の形で微分できないため図 3.8[6]のように $\varepsilon \sim N(0, I)$ でノイズを発生させて $z = \mu(X) + \varepsilon * \sigma(X)$ に変換する. これにより end-to-end で誤差逆伝播法が適用できるようになった.

VAE のケーススタディ

VAE の潜在変数はガウス分布で近似した確率分布で表せるため, 図 3.9[7]のように近い要素が集まる傾向になる. この例では学習データにない 9 と 4 の中間の数字画像も生成できている. これを言語に応用することで学習データを直接用意してなくても, 新たなサンプルを生成できる可能性がある.

次の章では 3 章の内容を踏まえて, 発送支援システムを実装するための提案手法を紹介する.

[6]https://qiita.com/kenmatsu4/items/b029d697e9995d93aa24
[7]https://arxiv.org/pdf/1312.6114.pdf

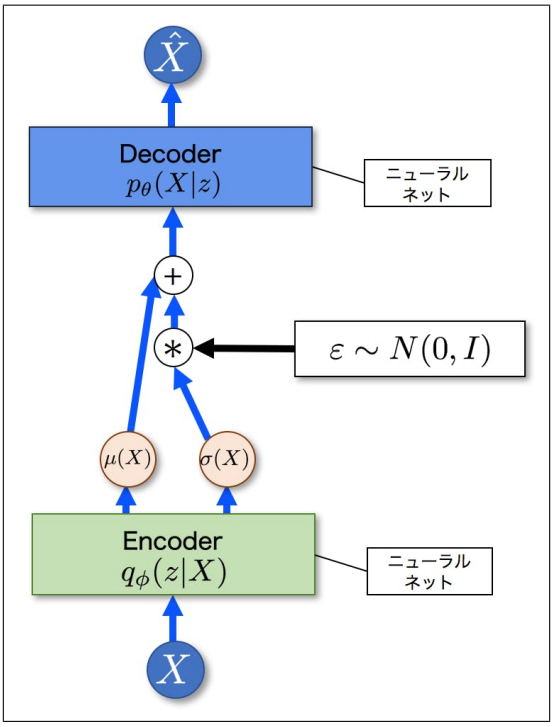
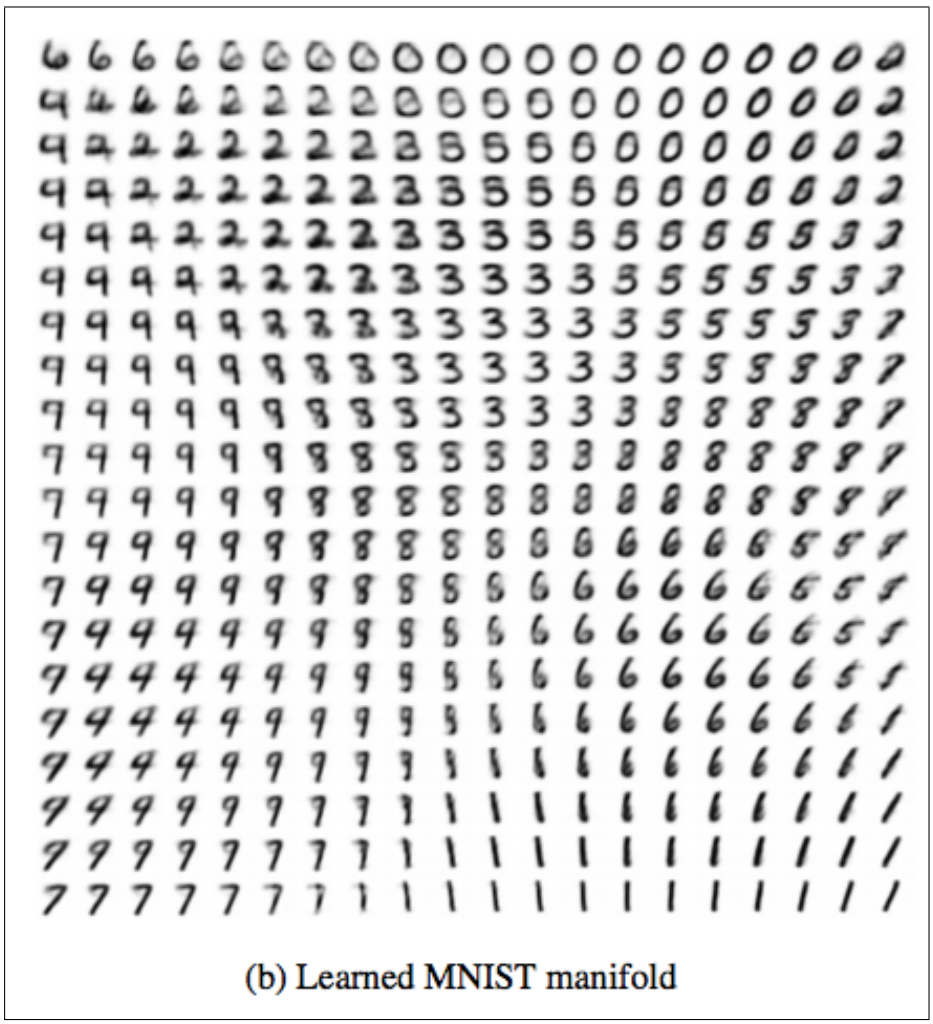


図 3.8: VAE のアーキテクチャ



(b) Learned MNIST manifold

図 3.9: VAE を用いた数字画像生成

第4章

提案手法とシステムのアーキテクチャ

§ 4.1　テキストデータのための VAE

前章では系列データを生成する seq2seq と LSTM, そして確率分布を推定する VAE を紹介した.

seq2seq を応用した生成系研究前章では翻訳タスクや対話等あくまで答えが用意されておりそれを予想できれば良いタスクに対してのアプローチを紹介してきた.

前述の通り本研究では答えとなる知財案を用意することが困難なため新しい文章を生成する必要がある. そのような先行研究としてターゲットの固有な特徴を捉えたキャッチコピーを生成する研究や画像をモチーフに俳句を自動生成する研究がある [32][33].

本研究のため仮説本研究の目的は文章を生成することだけでなく, 特許同士の関係性や発展性を考慮した出力が必要であるため上述の研究と同じ手法は適用できない.

そこで, 特許のテキストデータを意味のあるベクトル空間にマッピングできれば, 特許記事同士の演算や類似特許の検索、クラスタリング等様々の機能を内包したシステムが開発できると考えた.

従来でも単語レベルでのベクトル化ではすでに word2vec や glove, そして文書のベクトル化では Doc2vec 等が存在する. しかし, これらはテキストデータ限定の手法であり, 複数のモダリティに対応していない.

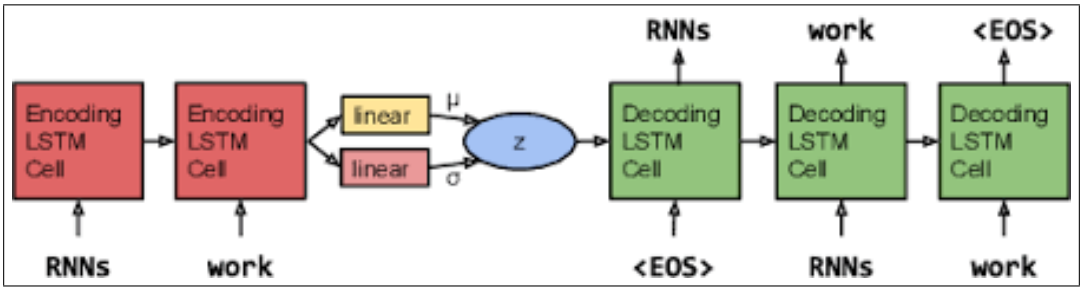


図 4.1: 文章のための VAE のアーキテクチャ

また, VAE はエンコーダ・デコーダにフィードフォワードニューラルネットワークを用いているため, 可変長の言語には利用できない. そこで Bowman らが文章のための VAE を提案している. 本稿ではこのモデルを Sentence-VAE と呼称する.

図 4.1 は Bowman らのモデルのアーキテクチャである. このアーキテクチャは VAE のニューラルネットを単純に LSTM に変えただけであるが, そのままでは学習がうまくいか

ないことが判明している.

そこで Bowman らは KL ダイバージェンスだけ学習する傾向があることを発見して学習を制御する KL weight なる係数を設け, 学習初期にはその係数 0 にすることで効果的に学習できることを示した.

また, デコーダだけ過学習する傾向があるため, デコーダの入力をランダムで単語を数値化する際に変換辞書に登録していない unknown word を示す＜ unk ＞ベクトルに置き換える **WordDropOut** という機構を組み込んだ.

図 4.2: Sentence VAE の出力例

これにより図 4.2 のように 2 つの文章間の補完する連続的な文章を生成できるようになった. 本研究ではこの Bowman の Sentence-VAE をベースとして考えていく.

§ 4.2　提案手法

特許データは特許の文書以外にも引用数・被引用数・発明者等のパラメータが存在する. 筆者はこのパラメータを有効活用することでより特許の特性を捉えたモデルが作成可能であると考えた.

NLP がメインのタスクで他のモダリティのデータを利用する場合いくつか方法が考えられる

1. 新しい入力分だけ入力次元を増やす
2. エンコーダに Multimodal learning を利用する [35]
3. デコーダの単語生成時にパラメータを考慮した重みを利用する
4. 損失関数にてテキスト以外の情報を報酬・罰則項に加える [36]

方法 1 はセンサデータをニューラルネットワークで学習する際に取られる方法である. センサが増えた場合, 単純にそれを新たな入力としてニューロン数を増やすだけである. しかし, 本研究の場合可変長でスパースなテキストデータに数次元の新データを追加するというネットワークになってしまう. そのためこの方法は利用できない.

次に方法は 2 は鈴木らが提案したモデルで MVAE と名付けられている [35]. 構成としては図 4.3 のようになりモダリティごとにエンコーダ・デコーダを用意して潜在変数である確率分布の平均と分散を 2 つのモダリティで共有することでマルチモーダルデータを処理することに成功している.

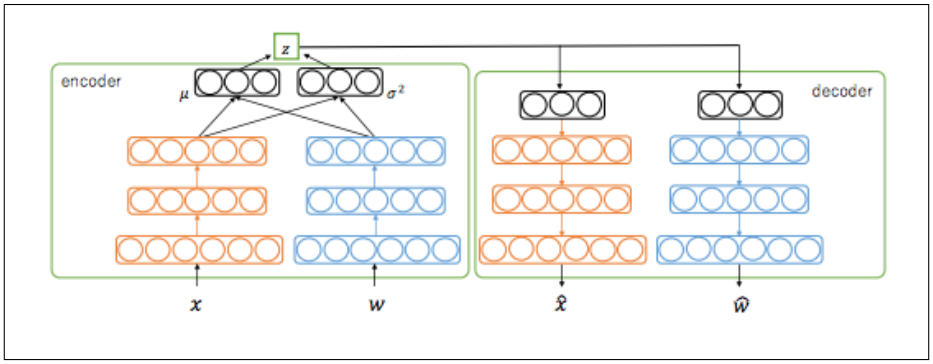


図 4.3: MVAE のネットワーク構造

しかし引用件数等の特許パラメータで収集可能なものは少なくその少ないデータでニューラルネットを構成する必要がある．そのため次元圧縮する VAE の利点が余り活かせない．

方法 3 は生成時に例えば Onoda が考案した DEA を用いて特許の引用数から単語の重み求める手法を利用できる [11]．これを利用することで, 引用数が高くなるような単語が選択されやすくなると考えられる．しかし本研究で行いたいのはテキストと特許パラメータどちらも反映したマッピングを得ることのためこの手法では本質的解決にならない．

最後の手法 4 の報酬・罰則項に特許パラメータを加えるという方法はニューラルネットを用いた学習でよく使われる方法論である．Bowman の手法では損失としてクロスエントロピー誤差と KL ダイバージェンスを利用していた．

ここに特許の引用数を考慮した報酬項を追加することで, 引用数を考慮しつつも end-to-end で学習可能なモデルが実装できる．

§ 4.3　特許データクローラーの作成

特許のオープンデータとしては NTCIR というカンファレンスのタスクを評価するための NTCIR-6 patent[1] テストコレクションが配布されているが申請が必要な点と引用数等パラメータが十分でないため今回は独自に収集してデータベースを構築することにした．

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自体は patents.google.com ドメインで公開されている．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

[1] http://research.nii.ac.jp/ntcir/permission/ntcir-6/perm-ja-PATENT.html
[2] https://www.google.co.jp/?tbm=pts
[3] https://patents.google.com

- Google を情報収集のプラットフォームとして利用した.

特許取得には国際連合の専門機関である World Intellectual Property Organisation が 2008 年発表したレポートにある特許分類を用いて英語の特許を収集した. 収集した分野のリストを表 4.1 に掲載する.

収集するデータとしては以下の 8 つである.

言語の特徴

- 特許 ID
- 特許タイトル
- 発明者
- 承認日
- 要約
- 本文
- 引用特許数
- 被引用特許数

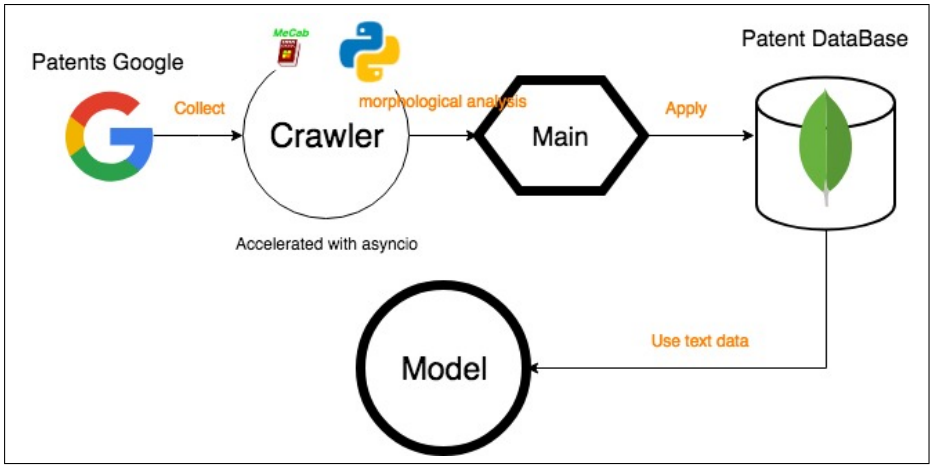


図 4.4: 特許クローラーのエコシステム

上記の収集対象のデータを蓄積・分析するために図 4.4 のような構成のデータベース (DB) 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. システムとして日本語でも英語でも収集できるようにした. 具体的には日本語で収集する際は

形態素解析を行うが, 英語の際ははじめから単語が分割されているためそのままそのまDB保存する

§ 4.4　データの前処理

まず特許データの内, 特許の価値として最も重要な指標として引用数・被引用数が挙げられる [3][37]. また他のデータとして論文引用数, 論文被引用等があるがこれは筆者が利用しているプラットフォームでは事実上収集できない. また, 承認日の情報は出願中の特許では勿論存在しない.

そこで出願中・承認後, どの特許でも収集可能な指標を利用することにした.

特許のテキストデータ

表 4.1: 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

表 4.2: テストモデルの学習結果

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

特許データのテキストデータは要約, 明細部, 請求項の 3 つに分けられる. すべて使うべきかは研究によって異なる. はじめに, すべての部分に対して以下の表 4.2, 4.3 の設定でテストモデルを作成して検証した. その結果以下のようになった. なお入力データに森らが考案した専門用語抽出システムを適用した特許重要キーワードを利用した [38]. 出力をそのまま特許全文を対象とした.

seq2seq による生成結果（上手く行った例）

-
- 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
- 予測結果
  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

このようにある程度正解の単語を予測できていたが, LSTM の特徴として, 今回のように長文の予測になると以下の “power” ように同じ単語を出力しやすくなる.

seq2seq による生成結果（失敗例）

- 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
- 予測結果
  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 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

以上の結果から学習対象は長文では効果を発揮しづらく, また本研究は長文の予測精度上昇に対して焦点をおいてないため今回は要約部のみを利用した.

第5章

数値実験と考察

第6章

おわりに

本研究の目的は，多くの人に広く受け入れられるライフログとして，個人情報保護に着目し，手間がかからず自動的にライフログデータの取得を行い，取得したデータから類似性やイベント性を考察できることである．開発したライフログデータ取得アプリケーションを使用したビッグデータ構築・データ解析を行い，行動パターンの類似性・イベント検出を行った．

結論として，個人情報保護に着目したライフログデータ取得アプリケーションの開発ができ，多変量解析を用いることでライフログの可視化を行い行動パターンの類似性やイベント性を視覚的に検出するという目標は達成できた．特に，SOM の解析結果より，同じ行動でも視界に写る物体の違いから行動の類似性やイベント性を検出できた．同じ行動でも使用する場所や物体の変化によって別行動として認識させることができるため，ライフログデータに位置情報を付加できると考えられる．よって，個人情報保護に着目し取得したライフログデータから類似性やイベント性を検出できたと考える．本研究の研究成果は，テキストによるライフログデータ取得，解析を行い新たなビジネスプランの検討やユーザー自身の生活の見直しなどに使用できるため，より高度なアプリケーション開発を目指す開発者，研究者の方々の参考になれば幸いである．解明できた点は必ずしも多くはないが，若干なりとも寄与できたと思われる．

今後の課題として，開発したアプリケーションの改善点を上げる．開発したアプリケーションは自動的にライフログデータを取得する点が利点として挙げられるが，一方で客観的なライフログデータしか取得できないという弱点もある．ユーザーが興味を持った瞬間や，データを取得したい瞬間のライフログデータは現状のアプリケーションには含まれていないためである．この弱点に対し，ユーザーが取得したいタイミングでライフログデータを取得する方法をアプリケーションに組み込む必要がある．組み込むため，取得したいタイミングを MOVERIO^{TM} に伝える方法の検討も必要となる．

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学電子・情報工学科の奥原浩之教授に深甚な謝意を表します．最後になりましたが，多大な協力をして頂いた，奥原研究室の同輩諸氏に感謝致します．

2020 年 2 月

小野田 成晃

参考文献

[1] 総務省, “総務省 ICT 利活用の促進 地方公共団体のオープンデータの推進”, https://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/, 閲覧日 2018, 5, 5.

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

[3] 藤井敦, 山川英和, 岩山真, 難波英嗣, 山本幹雄, 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.

[4] H. Mase, et al, Pro-posal of Two-Stage Patent Retrieval Method Considering the Claim Structure, ACM *Transactions on Asian Language Information Processing*, 4, 2, pp. 186 202, 2005.

[5] T. Kiriyama, T. Ando, “Patent Information and AI: Outline”, *Information Science and Technology*, Vol. 67, No. 7, pp. 340-349, 2017.

[6] Young Gil Kim, Jong Hwan Suh, Sang Chan Park, ”Visualization of patent analysis for emerging technology”, *Expert Systems with Applications*, vol. 34 pp. 1804-1812, 2008.

[7] 大田貴久, “機械学習等の情報技術を用いた特許調査について”, 情報の科学と技術, 67(7), pp. 366-371, 2017.

[8] 酒井浩之, 野中尋史, 増山繁, “特許明細書からの技術課題情報抽出”, 人工知能学会論文誌, 24 巻, 6 号, *I*, 2009.

[9] 津村拓海, 斎藤史哲, 石津昌平, “ランダムフォレストを用いた特許に関する文書データからの技術適用領域に関する知識抽出”, 日本経営工学会論文誌, vol. 68, No. 3, pp. 161-170, 2017.

[10] Hyonju Seol, Sungjoo Lee, Chulhyun Kim, “Identifying new business areas using patent informatin: A DEA and text mining approac”, *Expert Systems with Applications*, vol. 38, pp. 2933-2941, 2011.

[11] S. Onoda, K. Okuhara, “Selection of Core words from Textual Patent Data with DEA based on Citation”, 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 175-180, 2019.

[12] 川喜田二郎, “発想法 改版 創造性開発のために”, 中公新書, 2017.

[13] “KJ 法を支援するソフトウェアについての研究”, 皆川優也, http://www.net.c.dendai.ac.jp/ minagawa/, 閲覧日 2020, 1, 4.

[14] 伊藤淳子, 東孝行, 宗森純, “単語共起度の低い単語を提示する発送支援システムの提案と適用”, 情報処理学会論文誌, vol. 56, No. 6, pp. 1528-1540, 2015.

[15] 西原陽子, 日比野純也, 福本淳一, 山西良典, “希少な機能の提示による新商品のアイデア発想支援システム”, 知能と情報, vol.27, No.4, pp. 669-679, 2015.

[16] 黒橋禎夫, “自然言語処理”, 一般社団法人 放送大学教育振興会, 2015.

[17] R. Eric. “Jargon File Version 2.8.1”, http://catb.org/esr/jargon/oldversions/jarg282.txt, 閲覧日 2020, 1, 5.

[18] J. Devlin, M. Chang, K. Lee, K. Toutanova “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *In North American Association for Computational Linguistics (NAACL)*, 2019.

[19] 高村大地, “言語処理のための機械学習入門”, コロナ社, 2010.

[20] J. L. Elman “Finding structure in time.”, *Cognitive Science*, 14, pp. 179-211, 1990.

[21] SAS Institute Japan 株式会社, “ディープラーニングは、時系列予測でも最強なのか？ ～ RNN と従来手法との対比から見える使いどころ～”, https://www.sas.com/content/dam/SAS/documents/marketing-whitepapers-ebooks/sas-whitepapers/ja/viya-recurrent-neural-network.pdf, 閲覧日 2019, 12, 07.

[22] I. Sutskever, O. Vinyals, Q. V. Le, “Sequence to Sequence Learning with Neural Networks”, *In Advances in Neural Information Processing Systems (NIPS 2014)*, 2014.

[23] 松井藤五郎, 汐月智哉, “LSTM を用いた株価変動予測”, 人工知能学会全国大会（第 31 回）, 2017.

[24] 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.

[25] D. Bahdanau, K. Cho, Y, Bengio, “Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio”, *ICLR*, 2015.

[26] C. Li, W. Xu, Si. Li, S. Gao, “Guiing Generation for Abstractive Text Summarization based on Key Information Guide Network”, *proceeedings of NAACL-HLT 2018*, pp. 55-60, 2018.

[27] L. Sha, et al, “Order-Planning Neural Text Generation From Structured Data”, *AAAI Conference on Artificial Intelligence*, pp. 5414-5421, 2018.

[28] R. Puduppully, L. Dong, M. Lapata, “Data-to-Text Generation with Content Slection and Planning”, *AAAI Conference on Artificial Intelligence*, 2018.

[29] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, “Show and Tell: A Neural Image Caption Generator”, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[30] G. E. Hinton; R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Networks”, *Science*, 313 (5786), pp. 504-507, 2006.

[31] D. P. Kingma, M. Welling, “Auto-Encoding Variational Bayes”, *ICLR*, 2014.

[32] 三沢翔太郎, 西埜徹, 谷口友紀, 三浦康秀, 佐藤政寛 大熊智子, “対象に固有な特徴を捉えたキャッチコピー生成”, 言語処理学会 第 25 回年次大会 発表論文集, pp. 1293-1296, 2019.

[33] 米田航紀, 横山想一郎, 山下倫央, 川村秀憲, “LSTM を用いた俳句自動生成器の開発”, *The 32nd Annual Conference of the Japanese Society for Artificial Intelligence*, 1B2-OS-11b-0, 2018.

[34] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R, Jozefowicz, S. Bengio, “Generating Sentences from a Continuous Space”, *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

[35] 鈴木雅大, 松尾豊 “深層生成モデルを用いたマルチモーダル学習”, *The 30th Annual Conference of the Japanese Society for Artificial Inteligence*, 2016.

[36] 高山隼矢, 荒瀬由紀, “自己相互情報量を用いた特徴語彙予測に基づく雑談応答生成”, 言語処理学会 第 25 回年次大会 発表論文集, 2019.

[37] 後藤晃, 玄場公規, 鈴木潤, 玉田俊平太, “重要特許の判別指標”, *RIETI Discussion Paper Series*, 06-J-018, 2018.

[38] “”専門用語（キーワード）自動抽出システム”のページへようこそ”, http://gensen.dl.itc.u-tokyo.ac.jp/, 閲覧日 2019, 10, 25.

付録

A. 1　Hello World を並列実行するソースコード

Raspberry Pi 3 を 8 台で Hello World を並列実行するソースコード A.1 を示す.

A. 2　円周率計算を並列分散処理するソースコード

Raspberry Pi 3 を 8 台で円周率計算を並列分散処理するソースコード A.2 を示す.