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

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., October 25, 2018,
Toyama Prefectural Univ.

1.1. 本研究の背景

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

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

[1] http://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

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

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...

2.1. 知財創造とは

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

2.1. 本研究が目指すシステム

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

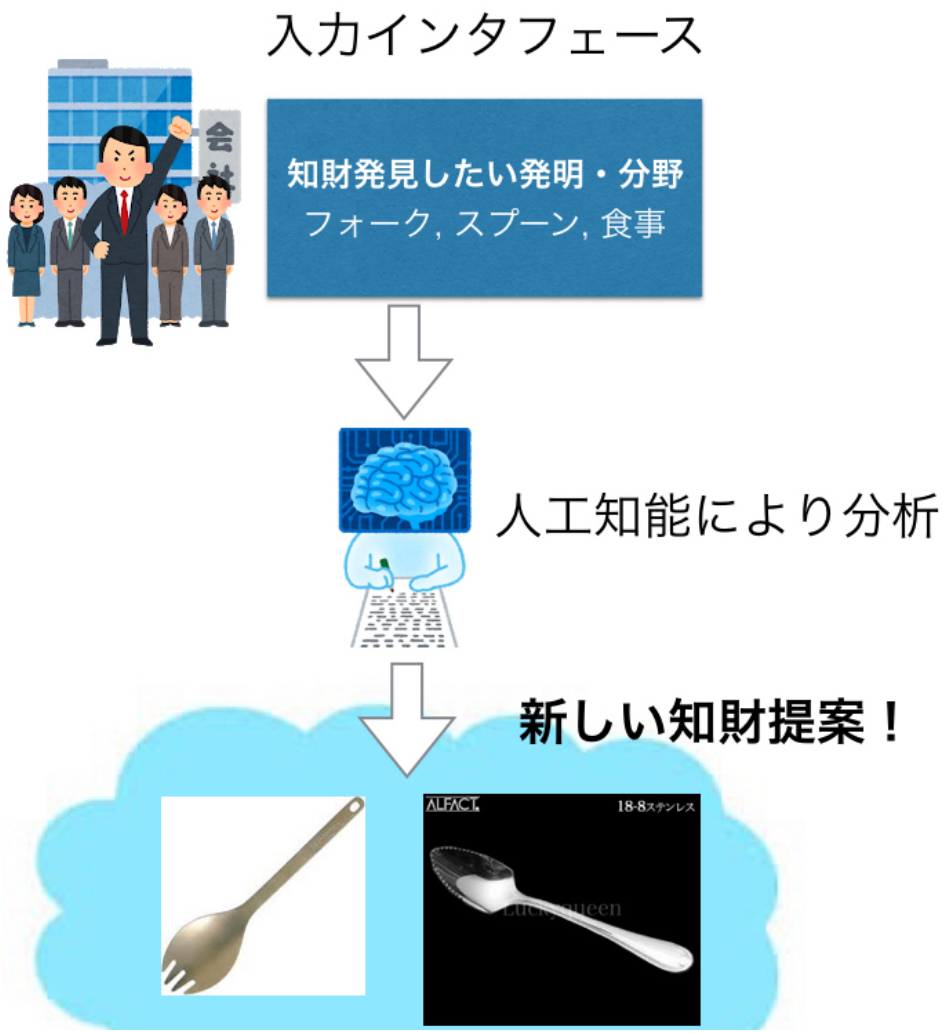

Figure: 1: システムの概念図

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.

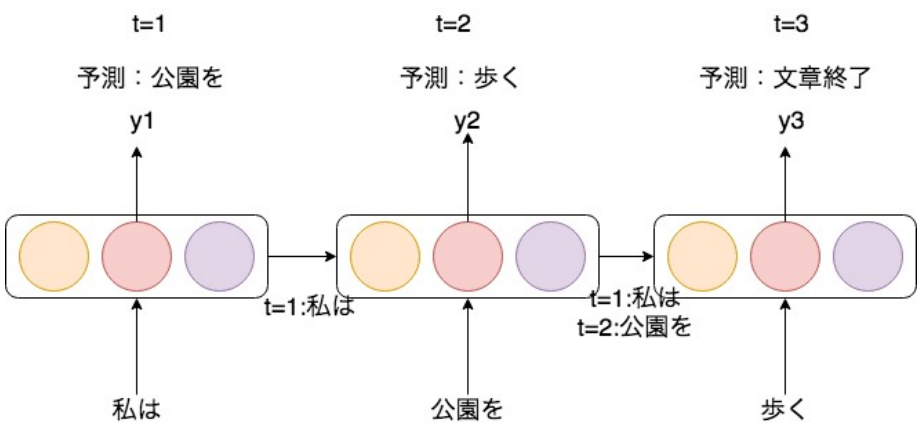

Figure: 2: 言語生成モデルの概念図

2.3. 生成モデルと LSTM

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

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 LSTM[3] を用いたものを使うのが主流.

[3] Long Short Term Memory

3.1. 系列変換モデル

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

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.

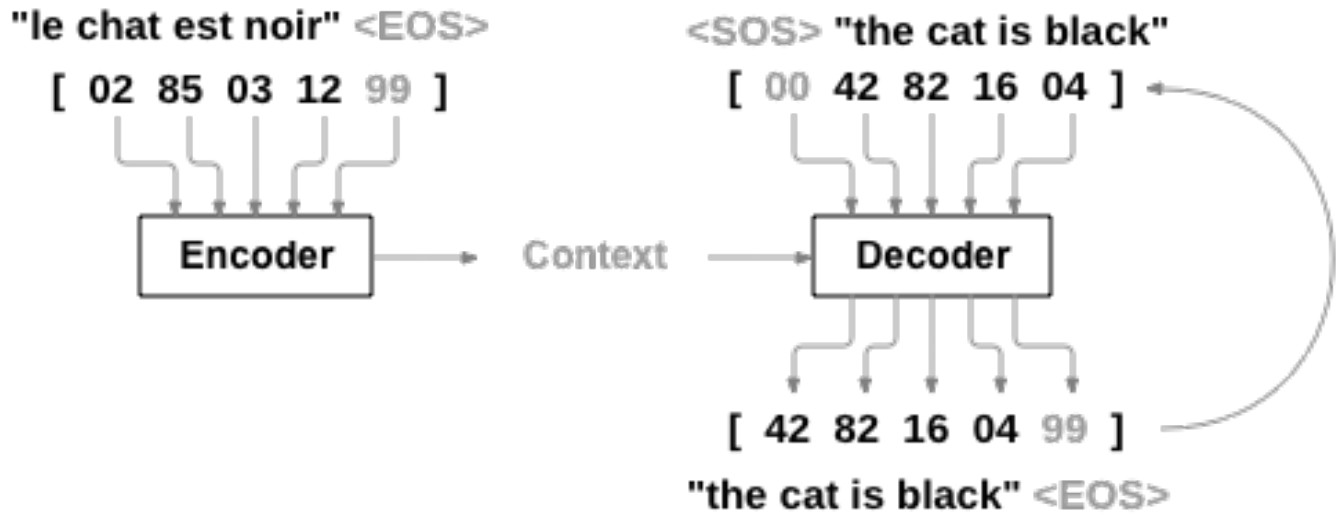

Figure: 3: フランス語から英語翻訳ネット

3.2. seq2seq の詳細

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

seq2seq の定式化

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

$x_1, ..., x_n$: 入力系列, $y_1, ..., y_n$: 出力系列, $\boldsymbol{s}$: 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.

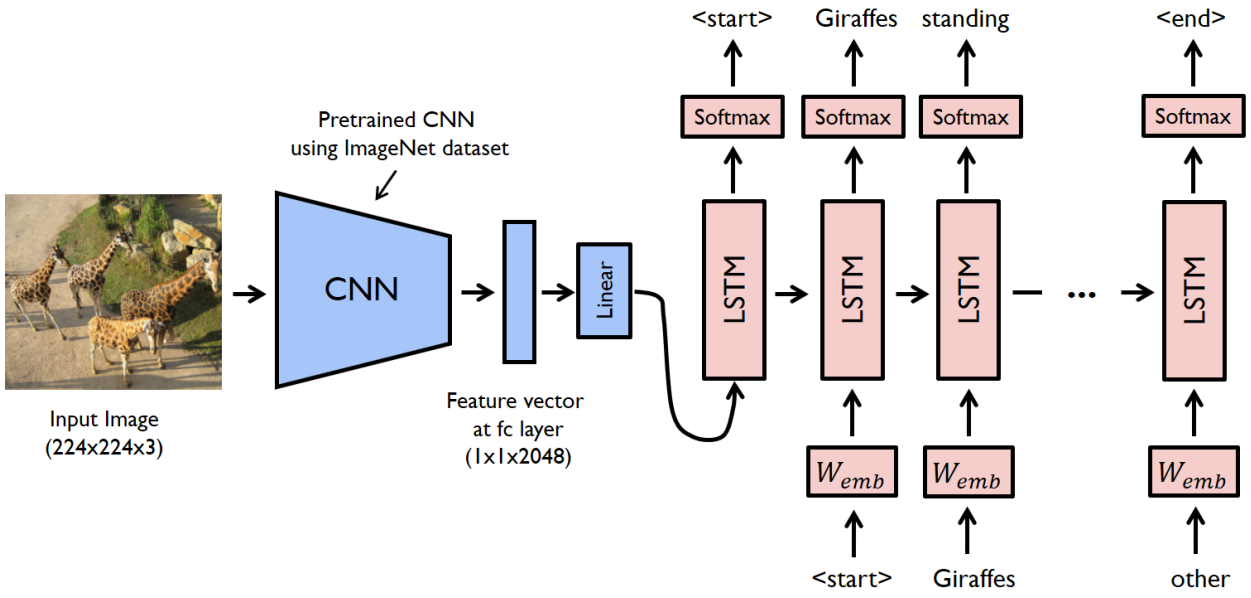

Figure: 4: 神経回路網による自動画像キャプション生成

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

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 Google**[4] 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

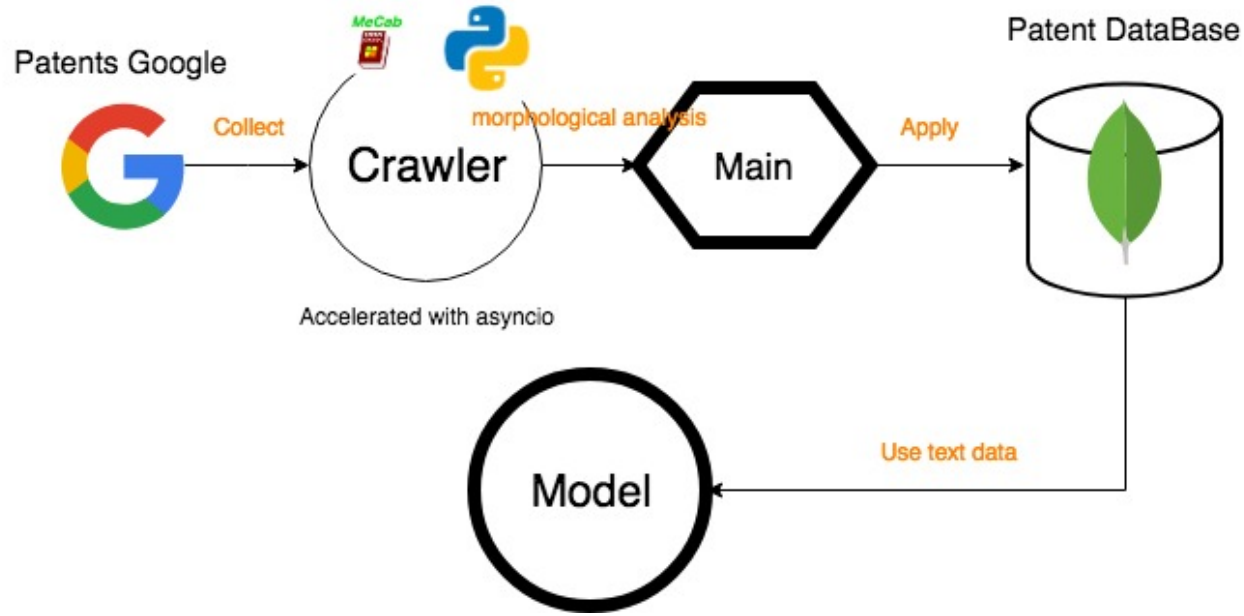

Figure: 5: 収集・分析プラットフォーム

4.2. 特許の価値の抽出

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

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

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

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.

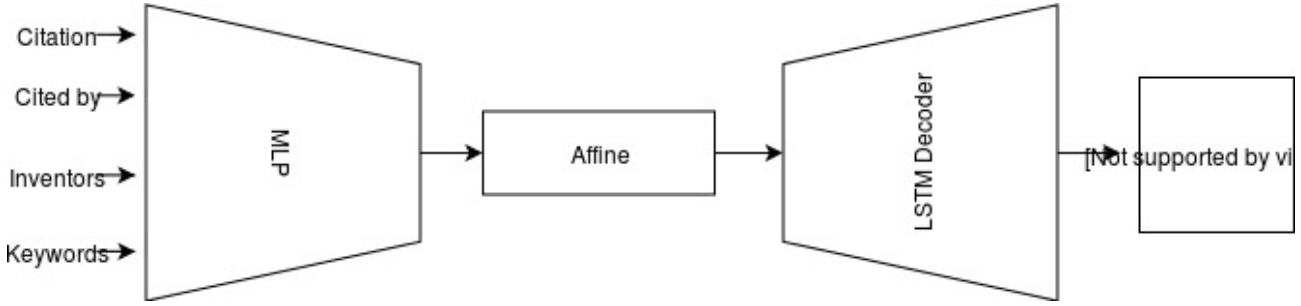

Figure: 6: 新しい提案モデル

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

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

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えるデータ包絡分析 (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

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

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

5.3 実験結果

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

サンプルのみだと。。。。

1. 29/04/12 IIS Hotel-Casino Hotel-Casino McSlarrow Hotel-Casino Hotel-Casino McSlarrow Hotel-Casino McSlarrow Hotel-Casino McSlarrow Hotel-Casino Hotel

当たり前であるが、このように一部の単語（Hotel-Casino）が確率分布から選ばれやすくなってしまう
=¿つまり学習が少ない場合はワンパターンな生成結果になりやすい

学習に一時間〜一日程度かかるため現在も学習中。。。

問題点

位置記事、約 5000 文字ほどあるので 10 文字程度の翻訳タスクより時間とリソースを食うことが判明した.
そのため, 長文処理の自然言語処理手法をサーベイする必要がある.

6. おわりに

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

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.
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

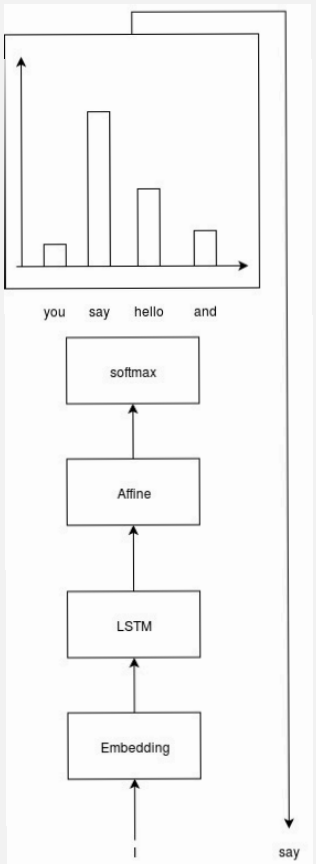

Figure: EX1: 生成モデル

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

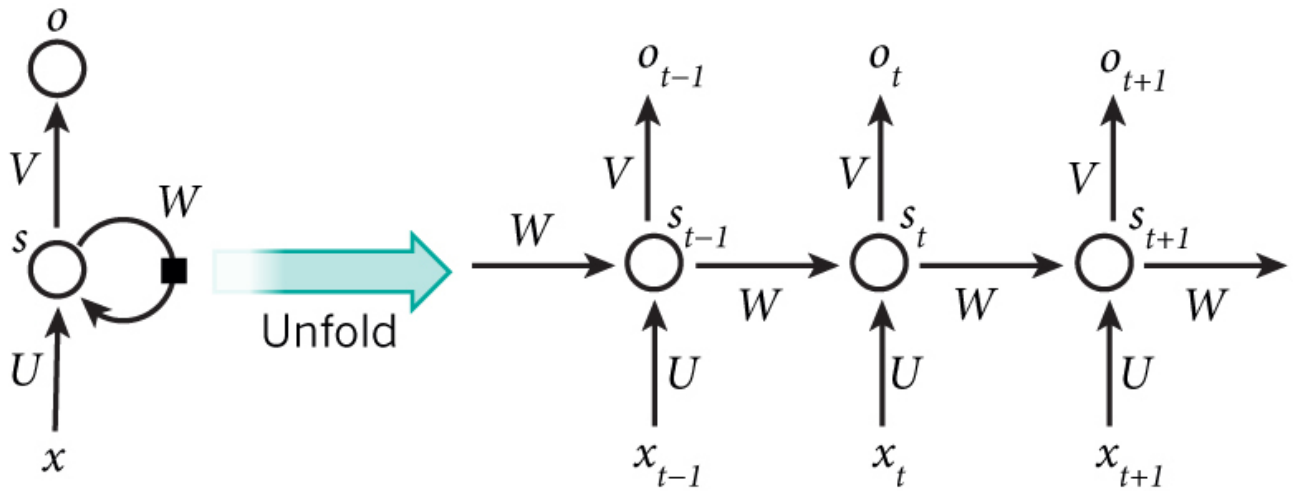

Figure: EX2: RNN

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

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

$$h_t = tahh(h_(t-1)W_h + x_tW_x + b)$$

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

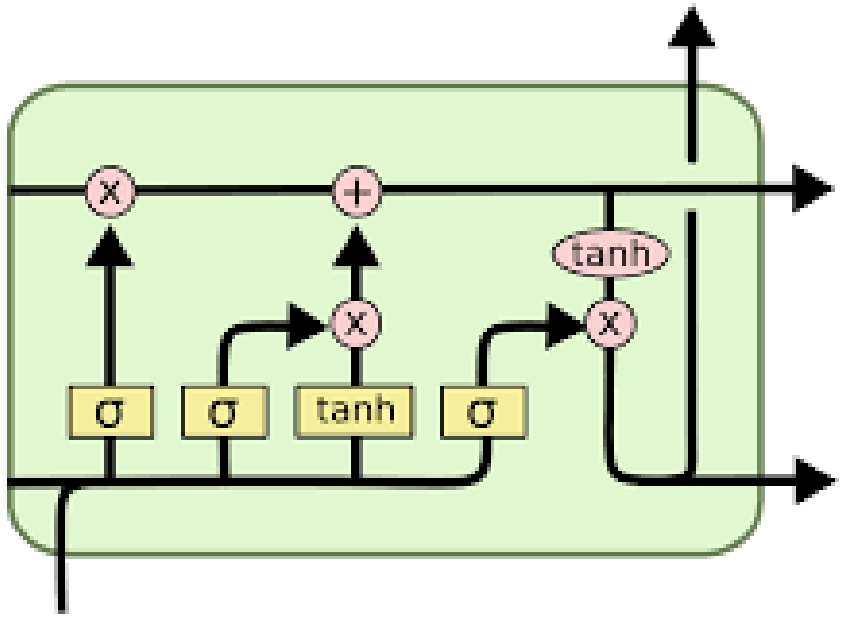

Figure: EX3: LSTM ゲートの概念図

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

LSTM

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

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

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

DEA

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

max :

$$y_k = \sum_{n=1}^{N} v_{kn} y_{kn}$$

subject to :

$$\sum_{m=1}^{M} u_{km} x_{sm} - \sum_{n=1}^{N} v_{kn} y_{sn} \geq 0$$

$$\sum_{m=1}^{M} u_{km} x_{km} = 1 \ (s = 1, 2, 3, ..., K)$$

$$u_{km} \geq 0 \ (m = 1, 2, 3, ..., M)$$

$$v_{kn} \geq 0 \ (n = 1, 2, 3, ..., N)$$

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

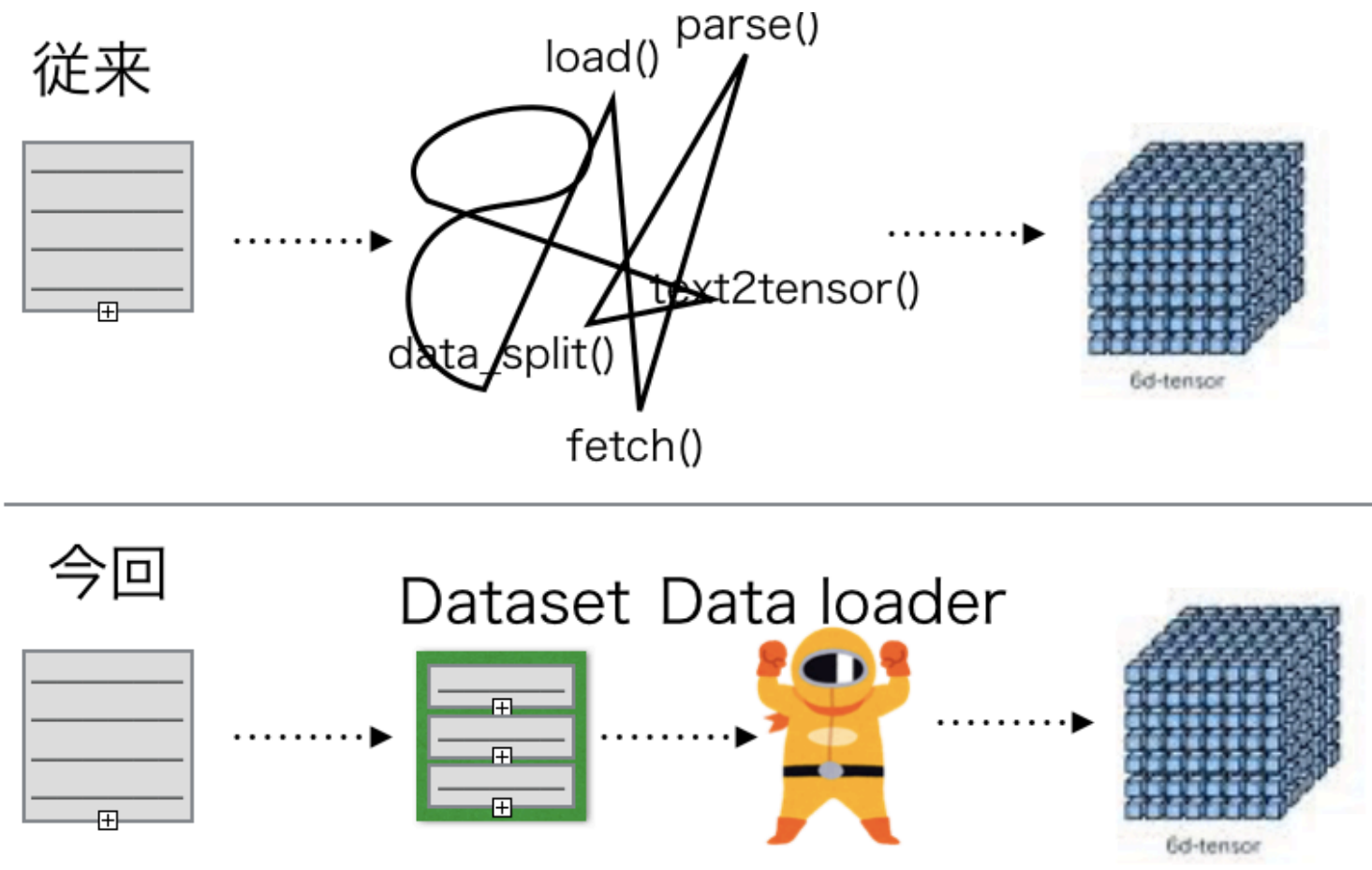

Figure: 1: データローダーの概要