特許情報収集による知的財産創造のための発見的の手法の開発

Shigeaki Onoda

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

情報基盤工学講座, 研究会

Friday., 11 November, 2018, Toyama Prefectural Univ.

アジェンダ

- 言語生成モデル
- seq2seq
- 特許情報エンコーダー

特許情報の現状と目的

現状

特許は知的財産の代表的なものである．しかし，綺麗に整理されておらず経営に活かされてるいるかは疑問符がつく．

目的

そこで本研究では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ことにした

特許データの複雑性

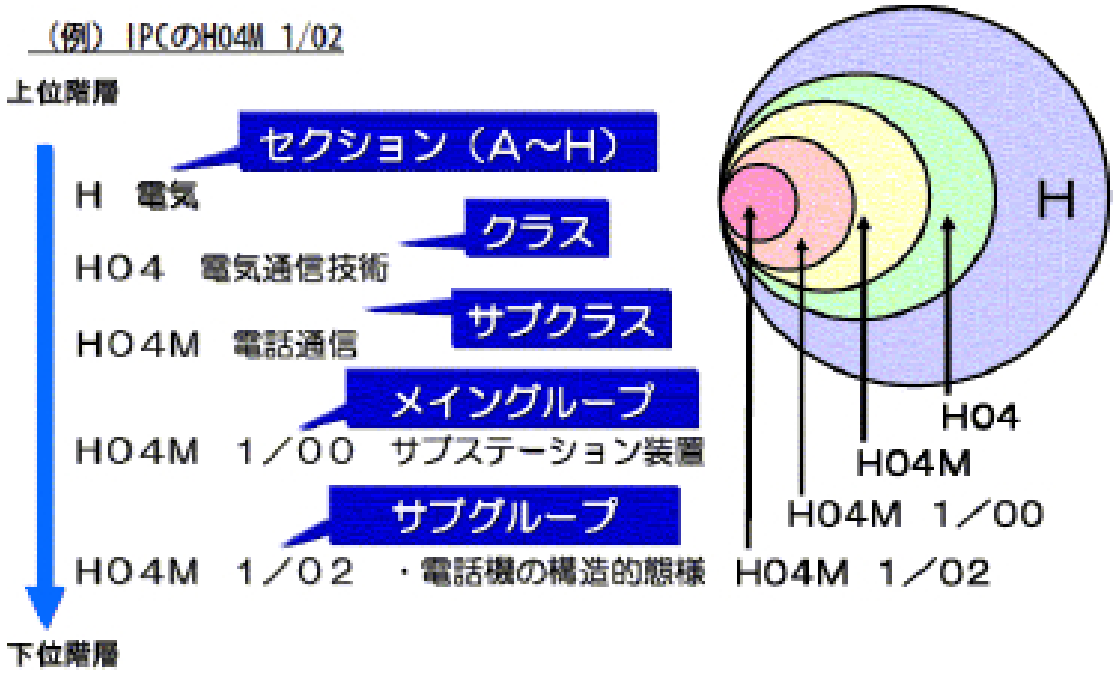


Figure: 1: 特許の分類図

このように一つの特許分野に対して深い小分類がある複雑なデータである

熊本県知財総合支援窓口 HP
http://chizai-portal.inpit.go.jp/madoguchi/kumamoto/consultation/support/patent/class/

先行研究

先行（関連）研究としては以下のものがある

1. ルールベースの重要特許判別指標
2. 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
3. 深層学習を用いたパテントマップ自動生成

本研究の枠組み

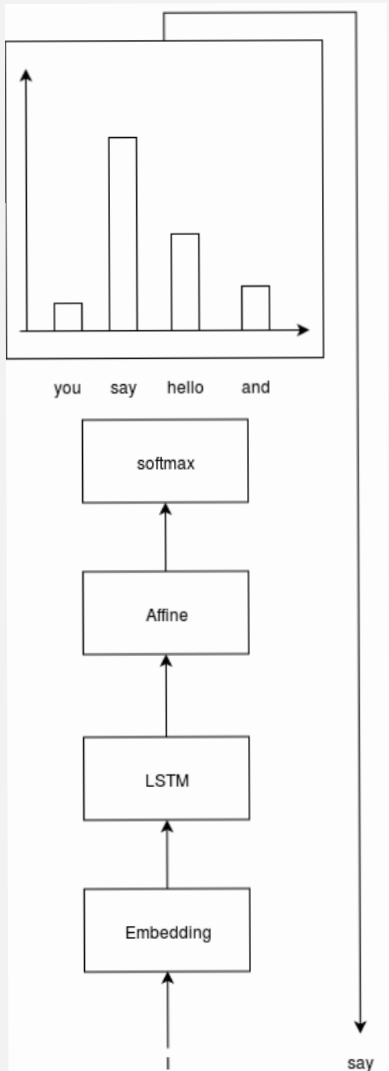
特許の文書と引用数等のパラメータを複合的に考慮している研究は少ない...

言語生成モデル

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



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

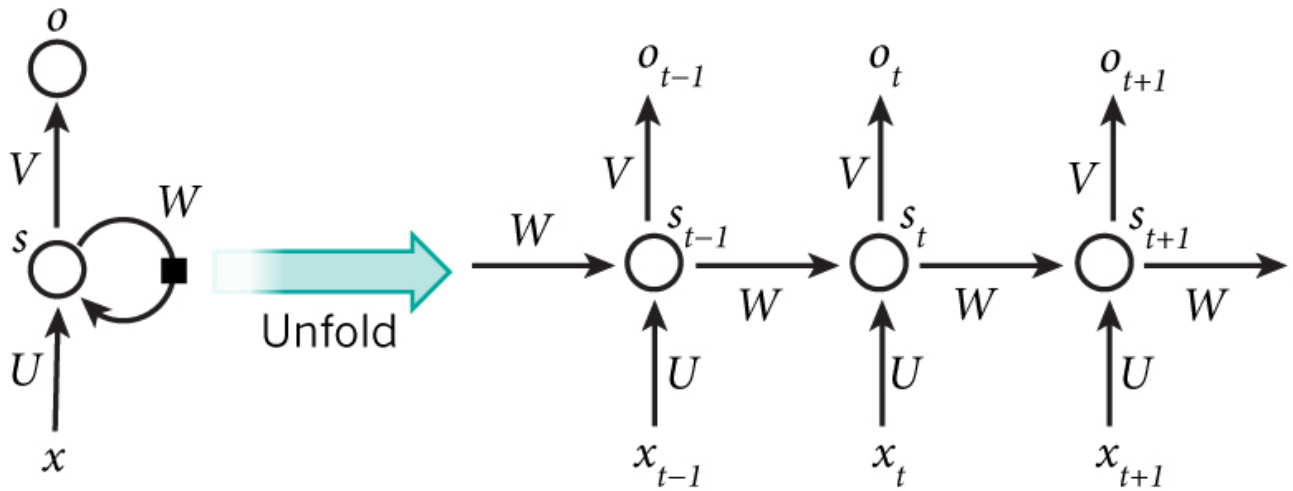


Figure: 2: RNN

LSTM

RNN では時系列の依存関係を学習できた しかし, かなり前の情報も前の情報も十把一絡げに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
ゲートと記憶セルという仕組みを導入してそれを改善したのが LSTM(Long Short Term Memory) である

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

seq2seq

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
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

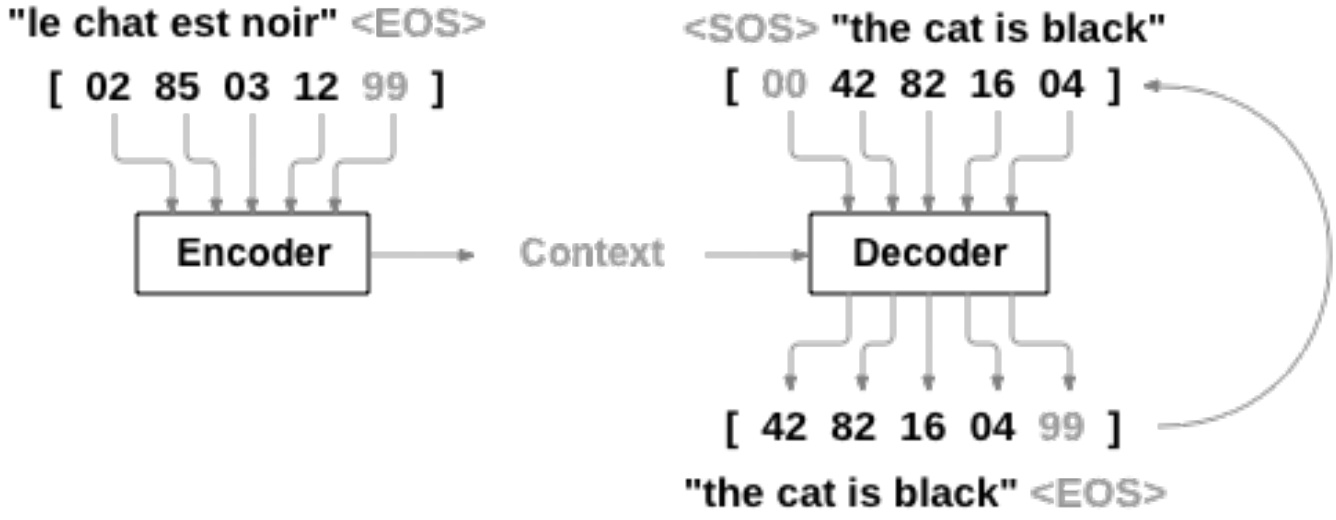


Figure: 3: フランス語から英語翻訳ネット

seq2seq の理論

seq2seq

seq2seq は全スライドの図のように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.

利点

大きな利点としては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.

提案手法

概要

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.
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

参考となるネットワーク構成

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なのでこのとき使われるノウハウを参考にして実装する予定

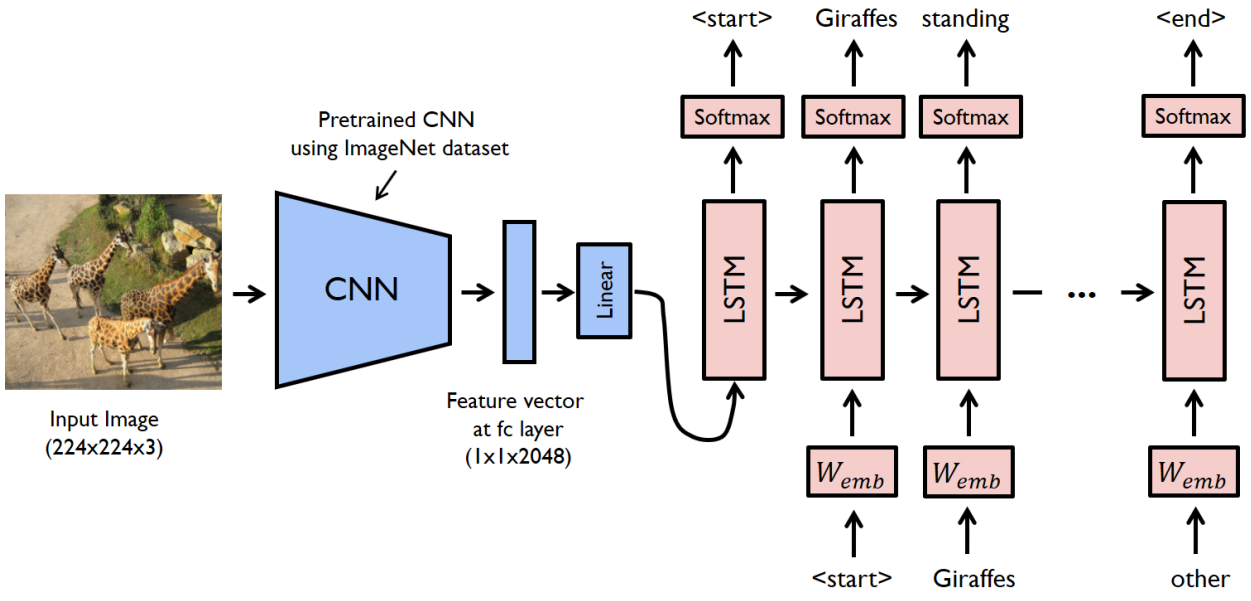


Figure: 4: A Neural Image Caption Generator

提案手法のコンセプト

行うこと

画像キャプションのモデルのように通常の 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られる

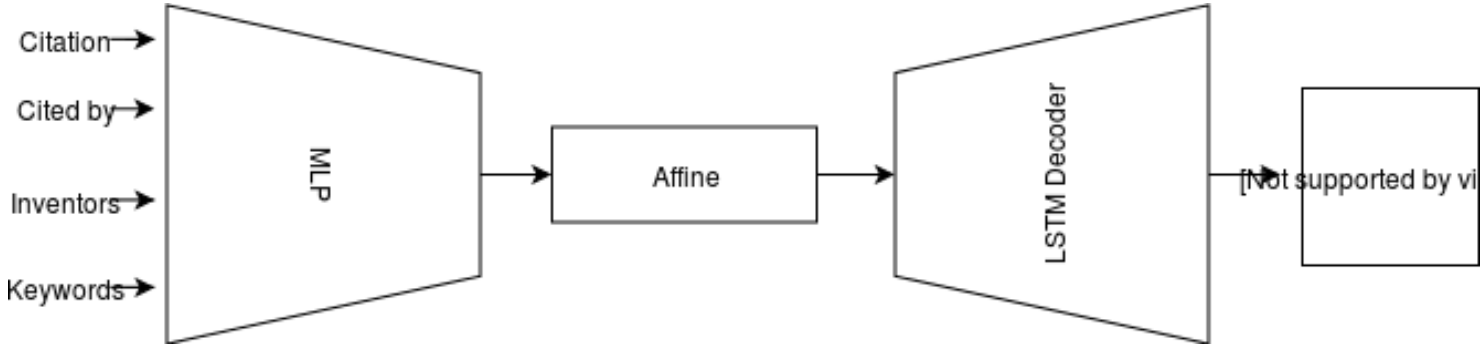


Figure: 5: 新しい提案モデル

コンジョイント分析との融合

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.

今後の展望・検討課題

概要

- 言語モデル作成時の素性を BoW, word2vec, fastText, glove 等マッチするものを検証
- 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