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

Wednesday., 9 18, 2019,
Toyama Prefectural Univ.

行ったこと

- ずっと詰まっていたところの解決
- プロトタイプモデルの実装
- 論文調査

やろうとしたこと

通常テキストマイニングではテキストを数値に表す方法として出現頻度のベクトルを使うがこれだと語彙数が増えるに連れて計算量も増えてしまうそこで計算量を節約するために固定長の入力ベクトルを使うことにした.

問題

学習済みの固定長単語ベクトルを使うときのベクトルサイズをあわせることが難しくモデルの作成で上手く組み込めなかったそのため詰まっていた.

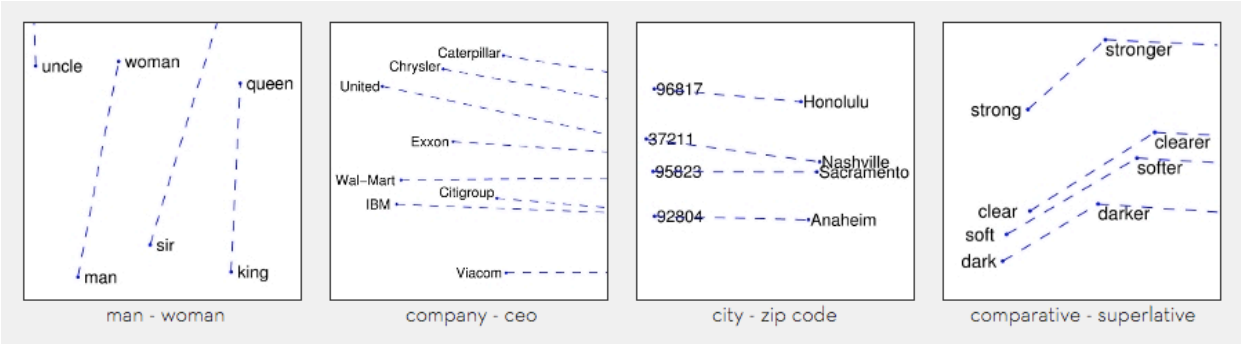

Figure: 1:glove モデルのイメージ

解決策

各ニューラルネット作成ライブラリでは埋め込み用のモジュールが実装されておりそれの使い方をきちんと学べば簡単であった

ソリューション

```
embedding = nn.Embedding(vocab_size, embedding_size)
embedding.weight = nn.Parameter(torch.from_numpy(weigths))
embedding.weight.requires_grad False
lookup_tensor = torch.tensor([word_to_ix[“hello”]], type=torch.long)
hello_embedding = embedding(lookup_tensor)
test_input = torch.Tensor([[19],[111],[837]])
embedding(test_input)
```

つまり ‘nn.Embedding ‘を定義した変数 ‘embedding‘に単語 ID を（複数でも）入力することで、word2vec 等の学習済み embedding ベクトルが得られる

プロトタイプ

翻訳モデルをベースとしてモデルのプロトタイプを完成させた構成は以下図 2

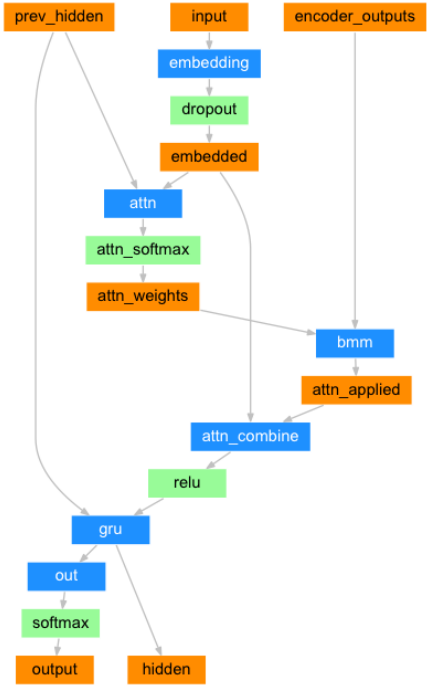

Figure: 2:プロトタイプの構成

構成

細かな設定項目

- 使用ネットワーク：LSTM, 注意機構
- 使用 word embedding：glove.840B.300d.model
web 記事から集めた 220 万語で 300 次元で構成される- 隠れユニット数：50
- 入力形式：キーワード
- 出力形式：文章（テストのため 10 文字以内）

現在の進度

小野田の想定

現在は超少量データでお試しの学習しているのみで本学習はしていない図 3 が現在のモデルから吐き出された混同行列 (Confusion matrix) である実データで学習してないため精度は良くないが一応モデルが動くことが確認できた

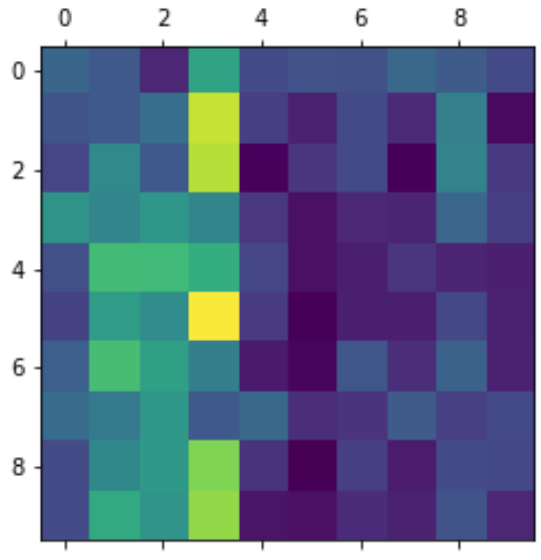

Figure: 3:結果

論文調査

開発と同時平行して行っていた論文調査で参考になりそうな論文を見つけたので紹介する

重要キーワードと記事からの要約文生成モデル

下の論文は自動要約のモデルであるが基本は小野田が行っている新規特許創案と同じ言語抽出・生成のため利用できると思われる.
なおキーワード抽出手法としてページランクの派生モデル TexRank を利用している＝＞山元君の研究に使える？

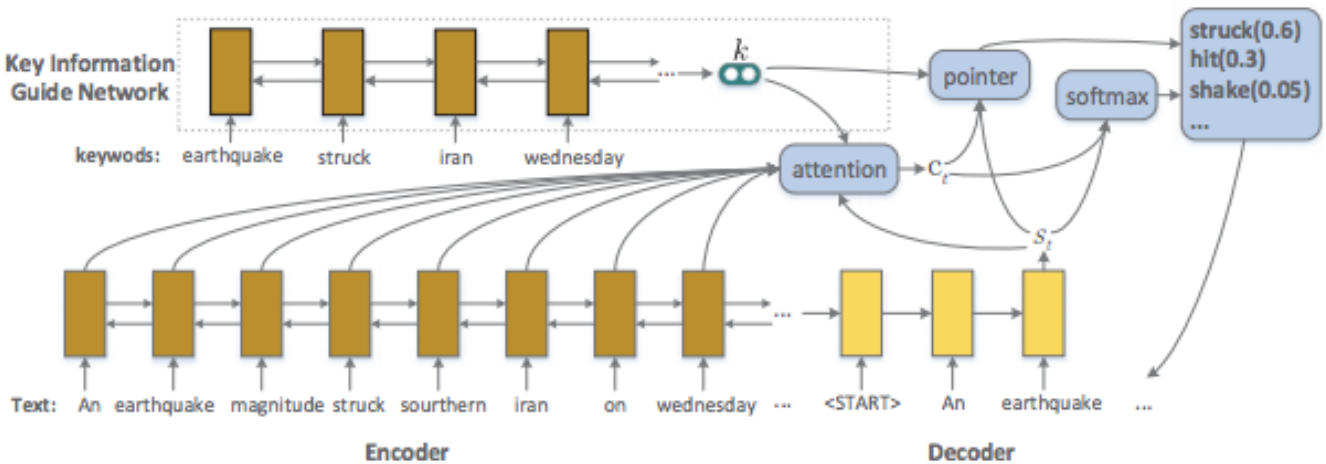

Figure: 4:ネットワークの概念図

URL: https://www.aclweb.org/anthology/N18-2009

今後やること

ToDo

1. 学習データの充実
2. 学習モデルの初学習
3. モデルのさらなる改良
4. フロントページの作成