WASEDA University

スマートエスイー 修了制作

Smart SE

社内開発技術QAチャットボットのプロトタイプ構築

株式会社クレスコ　奥村師範

背景と課題

社内では各開発分野の専門家を集めたQA対応チームが存在する。QA対応メンバーの主業務は別にあり、各プロジェクト業務中に各々対応している。また質問時は社内のRedmine上にチケットを起票しなければならず、質問者、回答者双方に負担である。

手法・ツールの適用による解決

もう少し気軽に質問できかつ回答者への負担を減らすため、QAチャットボットにより回答を自動化する場合を考える。その上で質問に対し適切な回答を得るため、機械学習の一手法であるDoc2Vecの適用を試みた。

文書の検索方法

手順

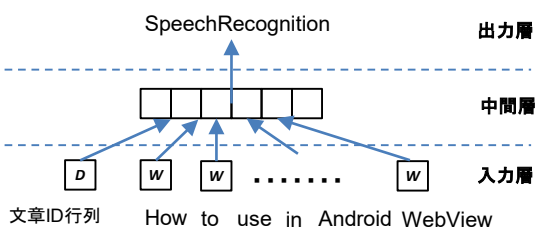
1. 文書のベクトル表現作成する。
2. 文章ベクトル同士のcos類似度を計算し、似た文章を割り出す。

従来手法

- **BOW**
  ⇒単語の出現頻度のみ
- **TF-IDF**
  ⇒単語の出現頻度×レア度

適用した手法

- **Doc2Vec (PV-DM)**
  ⇒中心語の出力確率が高くなるよう学習。
  中間層の重み行列を取得する。



- 語順が考慮できる。
- 単語ベクトルも一緒に学習できる。

Quoc Le, Tomas Mikolov. Distributed Representations of Sentences and Documents. Proceedings of The 31st International Conference on Machine Learning (ICML 2014), pp. 1188 – 1196, 2014

評価方法

1. stackoverflowのダンプデータを用意する。
2. android系タグでフィルタし、質問文を抽出したものをDoc2Vecで学習させる。
3. 学習したモデルに質問文を入力し、類似度の高い質問文に対応する回答文を出力する。

評価結果

- 学習に使用した質問文をそのまま入力した場合、該当文書のIDが返却され、期待した結果となった。
- 入力文を学習した文章より短くすると、全く異なる文書IDが返却され、期待した結果とならなかった。
- 単語の学習においては下記表のような結果となった。

表1　onCreateと類似度が高い単語ランキング

No.	単語	類似度
1	OnCreate	0.88406
2	onResume	0.81724
3	onDestroy	0.78496
4	onCreateView	0.78019
5	oncreate	0.77868
6	onStart	0.74779
7	setContentView	0.74639

表2　WebViewと類似度が高い単語ランキング

No.	単語	類似度
1	webView	0.89335
2	wv	0.79904
3	mWebView	0.78464
4	myWebView	0.77640
5	webview	0.77447
6	WebSettings	0.77128
7	AdvancedWebView	0.75880

考察

- 短い質問文に対応するためには？
  ⇒学習に使用した文章1セットは基本的に複数の文から構成されていたため、1文ごとに学習させてみるなどの改善案が考えられる。
- 学習前のデータ整形処理において、大文字/小文字の統一や不要な語や文章の削除、(～)の削除、変数の削除など改善すべき点がある。
- 単語の学習においては、開発に使用する単語において類似性が高い単語がランキングされることがわかった。学習データにコードを含んでいることが寄与していそう。