日本語 QA データセットを用いたクイズ解答システムの構築

情報科学科　浅野 綾太　　指導教員：小林 邦和

1　はじめに

近年、自然言語処理/質問応答の分野ではクイズ形式のデータセットが提案され注目されている。よく使われるものには SQuAD、NewsQA、TraviaQA などがあるが、これらはいずれも英語で作られている。日本語では「解答可能性付き読解データセット」や「JAQKET」[1] などが提案されているが、広く使われているわけではない。本研究では、日本語 QA データセット「JAQKET」を用いたクイズ解答システムの構築を目的としている。

2　データセット

本研究で用いるデータセット「JAQKET」は主に問題文 Q、答え A、20 個の解答候補 C_i、wikipedia を基に作られた解答候補の説明 P_i からできている [1]。

3　提案手法

本研究では、弱学習機 (モデル) を複数生成し、それらをアンサンブル学習の一つであるバギングの考え方を基に考案した方法を用いて正解率の高い解答システムを構築する。

3.1　弱学習機による解答の決定方法

ある問題に対し 20 個の解答候補それぞれに対して、その問題の解答である確率を算出し、その値が最も高かった解答候補をその問題の解答として出力する。「その問題の解答である確率」とは、問題文 Q と解答候補の説明 P_i の類似度としている。また、異なる回答候補で類似度が同じ値だった場合には、そのような解答候補の中からランダムで 1 つを選出するものとする。

3.2　モデル

問題文 Q と解答候補の説明 P_i の類似度を算出する弱学習機を生成する。

3.2.1　Simpson 係数を用いたモデル

Simpson 係数を用いて類似度を算出する。比較する文章は分かち書き又は N-gram を用いて配列に変換し、その配列を引数として Simpson 係数で類似度を得る。4.1 では、テキストを分かち書きして Simpson 係数で類似度を算出するモデルと、テキストを N-gram 処理をして Simpson 係数で類似度を算出するモデルの結果を示す。

3.2.2　コサイン類似度を用いたモデル

文章を配列に変換し、それぞれの単語をベクトルで表現し平均を算出する。それらをコサイン類似度を用いて類似度を算出する。ベクトル変換には TF-IDF と Word2vec を用いる。4.1 では、テキストを TF-IDF でベクトル表現し、コサイン類似度で累次度を算出するモデルと、テキストを Word2vec でベクトル表現し、コサイン類似度で累次度を算出するモデルの結果を示す。

3.3　複数の弱学習機を用いたバギング

図 1 に示す 3 つの方式を比較検討する。3.1 では、各モデルは類似度が最も高い解答候補を回答していた。方式 1 は 2 値分類問題を解くときのバギングの考え方である。方式 1 は類似度が同じだと解答がランダムになってしまうという欠点がある。方式 2 は回帰問題を解くときのバギングの考え方である。方式 2 では誤答選択肢でも類似度が同じように加算されてしまう欠点がある。方式 3 は方式 2 の欠点を補うために提案する方式である。方式 3 は順位におけるスコアの差が一定になっている点があるため、最適とは言えない。実験によって各方式の性能を検証する。

方式 1:	各モデルで類似度が最も高い解答で多数決し、解答候補を回答する
方式 2:	各モデルで算出された類似度の総和を算出し、累計類似度が最大の解答候補を回答する
方式 3:	各モデルで算出された類似度を順位に換算し、それぞれの累計順位が最小の解答候補を回答する

図 1　バギングの手法

4　計算機シミュレーション

4.1　各モデルの正解率

3.2 で示したモデルの正解率を表 1, 表 2 に示す。

表 1　Simpson 係数を用いたモデルのシミュレーション結果

ID	テキスト処理	平均正解率 [%]
⟨1⟩	無し	40.7
⟨2⟩	分かち書き	53.5
⟨3⟩	分かち書き (名詞のみ抽出)	60.3
⟨4⟩	N-gram(N=2)	56.9
⟨5⟩	N-gram(N=3)	62.3

表 2　cos 類似度を用いたモデルのシミュレーション結果

ID	テキスト処理	平均正解率 [%]
⟨6⟩	TF-IDF	37.6
⟨7⟩	TF-IDF(名詞のみ抽出)	43.8
⟨8⟩	Word2vec	17.5
⟨9⟩	Word2vec(名詞のみ抽出)	25.6

4.2　バギングを用いたモデルの正解率

4.1 で示したモデルのうち、⟨3⟩, ⟨4⟩, ⟨5⟩, ⟨7⟩, ⟨9⟩ を用いてバギングを行った結果を表 3 に示す。方式 1,2 はでは 4.1 で示したモデルよりもわずかではあるが高い正解率を得た。しかし、方式 3 では ⟨5⟩ の正解率を下回った。

表 3　バギングを用いたシミュレーション結果

学習方法	平均正解率 [%]
方式 1	64.4
方式 2	66.2
方式 3	57.1

5　おわりに

日本語 QA データセット「JAQKET」を用いてクイズ解答システムを構築した。今後の課題として、BERT などを用いたモデルを作成し、正解率の向上を図りたい。

参考文献

[1] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET:クイズを題材にした日本語 QA データセットの構築. 言語処理学会 第 26 回年次大会 , pp 237-240,2020.