

13点以下は取らせなAI

竜ヶ崎第一高等学校 税所理功 高山知幸 和田康佑

目的

1回のテストで45個の単語から15個が出題される英単語小テストで生徒の成績を上げるため、45個分の単語のデータからそれぞれの単語が出るか出ないかを予測する。

分類

学習データ...コーパス（英単語帳）の単語小テストに出題された単語（225単語）
単語を2つの特徴量（難しさ、品詞）に分け、数値化する。

* 単語の難しさはCEFRを参照

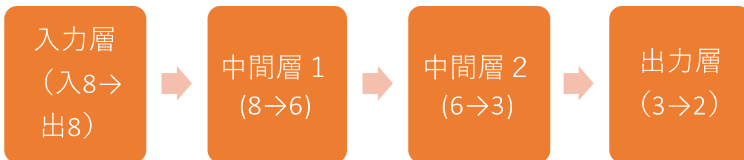
・数値の基準

難しさ (CEFR)	品詞
B2... 3	名詞... 0
B1... 2	動詞... 1
A2... 1	形容詞... 2
A1... 0	副詞... 3

これらの数値を、それぞれ4次元にone hot vectorへ変換し、リスト化したものを連結させた計8次元のリストを入力データとして扱った。

手法1(CNN)

入力データ...one hotに変換した8次元のリスト
出力データ...その単語が出題されたか(0,1)



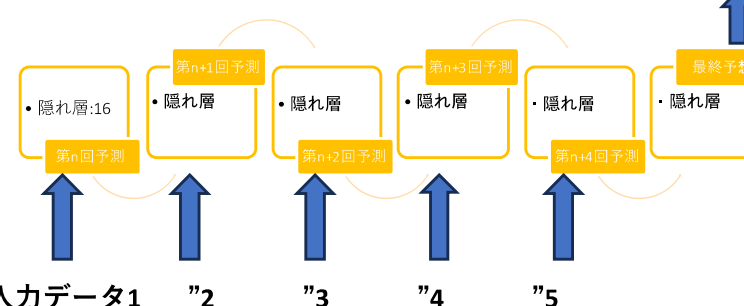
8つの入力データから出題（出るか出ないか）の2つのどちらかを予測する2クラス分類。
入力データ及び出力データはtensor型で32bitの少数にして扱う。

活性化関数はReLUを使用

手法2 (LSTM)

入力データ(×5)...one hotに変換した8次元のリスト

出力データ(×5)...その単語が出題されたか(0,1)



連続した単語テストのデータ5回分を時系列を考慮して学習させ、連続した5回のテストの次の回のテストの出題を予測。

学習1

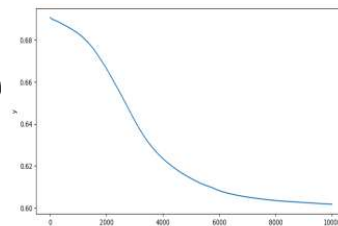
実際の英単語テストに出題された及び未出題の英単語、計225個分を用いて誤差を調べる。その後学習したモデルで小テスト一回分の範囲である、45個分の英単語を出るか出ないか分類した。

●学習用のデータの誤差の推移

学習回数...10000回

誤差計算...BCEloss(交差エントロピー誤差)

学習率...1e-3



・実際に予測した結果

	実際YES(15)	実際NO(30)
予測YES(6)	1	5
予測NO(39)	14	25

学習2

連続した実際の英単語テスト5回分に出題された及び未出題の英単語、計225(45×5)個分を用いて時系列を考慮（最初の回の予測(出力データ)を次の回の予測に反映させることを5回繰り返す）して学習し、5回のテストの次の回の出題を最終的に予測。

●学習のデータの誤差の推移

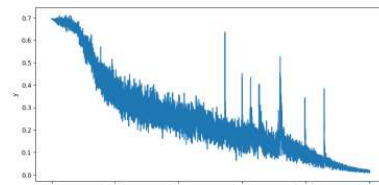
学習回数

...10000回

誤差計算

...BCEloss

学習率...1e-4



学習においてのモデルの精度ではRNNが上回っていたが実際の予測において精度があまりCNNとRNNで変わらなかった

・実際に予測した結果

	実際YES(15)	実際NO(30)
予測YES(23)	11	12
予測NO(22)	8	14

考察・課題

英単語テストにおいて時系列はあまり関係ないと考えられる。

・課題

学習に使うデータの量が圧倒的に足りないためデータが十分にある、他のテストや検定(英検など)で同じように試したい