クイズ問題のジャンル推定における機械学習手法の比較・検討

浅野　綾太 [1] 小林　邦和 [2]
[1] 愛知県立大学大学院 情報科学研究科 システム科学専攻 [2] 愛知県立大学 情報科学部
{im213001,kobayashi}@cis.aici-pu.ac.jp

概要

本研究では, 主にクイズ問題のジャンル推定を取り扱う. 先行研究として淀川らの研究 [1] を取り上げる. 淀川らは, 分散表現の獲得に MeCab と fastText を, 機械学習手法としてサポートベクタマシン (以下:SVM) を適用し, 問題文に対するジャンルを推定を行った. しかし, 先行研究では機械学習手法が SVM の１種類しか紹介されていなかった. そのため, 本研究では分類問題を解くその他の機械学習手法やグリッドサーチ, 交差検証を用いて最適な手法とパラメータを探索した.

結果としては, 単独では先行研究と同じく SVM によるジャンル推定が最高正解率を記録した.

1　はじめに

1.1　研究背景

クイズとは, 一方が問題を出題し, もう一方が答える「遊び」である. しかし, 問題の解答権を獲得することを競争することで競技性が生まれる. そのようなクイズのことを「競技クイズ」と呼ぶ. そして競技クイズは他の参加者に先んじて解答権を獲得するために知識を蓄える必要がある. 競技クイズの勉強法の一つに,Q 宅[1)]や BOOTH[2)]といったサイトで問題集を購入し, さまざまな問題に触れることで知識を得るという方法がある. このような勉強法の中でも,「芸能問題だけは正解できるようにしよう」というように, ある特定のジャンルに絞って問題を覚えるという方法がある. この場合, 芸能問題のみが掲載されている問題集を購入しなければならないが, そのような問題集は少数である. しかし, 市販の問題集の多くはジャンルが付与されていることは非常に少ない. 本研究では, 教師あり学習手法を用いてクイズの問題文からジャンルを推定する.

また, 本稿では便宜上「競技クイズ」のことを「クイズ」と表現する.

1.2　質問応答

ある質問に対する解答を求めるタスクを質問応答 (QA) タスクといい, しばしば実際に使われる問題が用いられることが多い. 自然言語処理における代表格として,「SQuAD[2]」「TraviaQA[3]」「QANTA[4]」などが挙げられるが, そのいずれも英語による文書を取り扱っている. 日本語のデータセットとしては,「解答可能性付き読解データセット [5]」や「JAQKET[6]」などがある. 実際のクイズ問題を扱った事例としては,2011 年 2 月に放映されたクイズ番組「Jeopardy!」で人間のチャンピオンを破った IBM の「Watson」などが挙げられる [7]. 早押しを想定した解答システムとしては, 橋元ら [8] がある.

1.3　先行研究

先行研究として, 淀川らの研究 [1] を取り上げる. 訓練データやテストデータには後述のデータセットを用いた. 分散表現の獲得には日本語 wikipwdia を用いた fastText を作成して利用した. モデルの訓練には, クイズの問題文を分かち書きし, 単語の分散表現を獲得し,300 次元からなる文章の分散表現を得る. 訓練データに 300 次元で表された問題文及び解答の解説文章と対応するジャンルを用いて SVM で学習する. 訓練データには以下の 2 種類を用いる.

(1)　「問題文」のみ
(2)　「問題文」と「解答の解説」

(1) では問題文のみを,(2) では, 問題文に対する解答を wikipedia の日本語データベースから説明文を獲得して問題文と結合し,(1) と同じくベクトル化・モデルの訓練を行う. 結果としては (1) で 40.0%,(2)54.8%をという正解率が得られた.

1）　https://q-tak.com/
2）　https://booth.pm/ja

2　利用するクイズデータセット

データセットには先行研究 [1] でも使われている「abc 13th 17th」と「EQIDEN 2013 2019」という大会の問題集を用いる[3)]. abc は学生を対象とする日本最大級のクイズ大会である. EQIDEN は abc と同日に行われる, 学業機関別のクイズ大会である. abc/EQIDEN で使用される問題は作成するにあたり多人数による審査が行われているため, 文章構成の質が良く, 難易度も比較的低く均等になっている. また, ジャンルにおいても偏りが少なくするように製作されている (図 1). そのため, 本研究で行うジャンル推定をするに当たり信頼性が高いと言える. 問題の用途は「早押し」「筆記・4 択」「敗者復活」に分類されるが, 本研究では「早押し」のみを使うこととする.

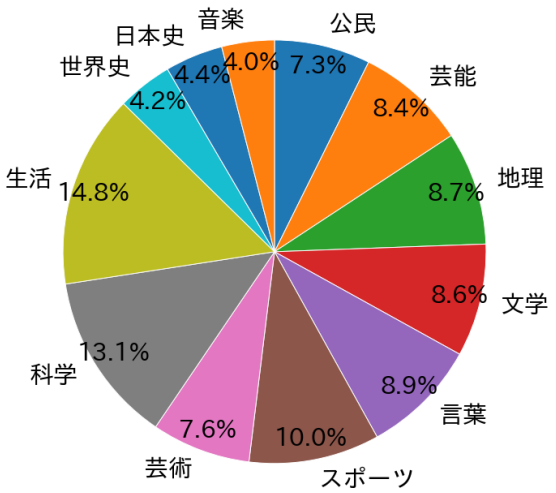

図 1　クイズ問題のジャンル比率

なお, 分類問題として取り扱う際に, 開催年度毎にジャンルが異なるため調整が必要になる. 13th から 15th は表 1 の 12 種類に,16th から 17th は表 2 の 15 種類に分類されている. 本研究では先行研究での 12 種類の分類 (表 3) するように微調整して利用する.

3)　https://abcdive.booth.pm/

表 1　abc13~15th のジャンル分類

文学	科学
地理	公民
生活	歴史
言葉	スポーツ
音楽	芸能
芸術	ノンセクション

表 2　abc16~17th のジャンル分類

自然科学	科学史
文学	思想・心理・社会学
言葉	日本史
世界史	地理
公民	芸術
漫画・アニメ・ゲーム	生活
スポーツ	芸能
ノンセクション	

表 4　用途別の問題数

大会	早押し	筆記	敗者復活
abc13th	800	150	15
EQIDEN2015	400	-	-
abc14th	800	150*	15
EQIDEN2016	320	-	-
abc15th	750	150*	15
EQIDEN2017	340	-	-
abc16th	800	200*	20
EQIDEN2018	400	-	-
abc17th	800	200	20
EQIDEN2019	320	-	-
total	5730	850	85

* ジャンルが付与されていない問題

3　実験方法

3.1　tokenizer

文字列を機械的に扱うとき, その文字列を分かち書きし, 分散表現を獲得する必要がある. 分かち書きには形態素解析エンジンの MeCab と新語・固有表現に強い辞書 NEologd を用いる.

分散表現の獲得には,Doc2Vec という手法を利用する. Doc2Vec は,Word2Vec の単語の分散表現を獲得する手法の応用である. なお, 本研究ではモデル作成の時間を省くために学習済みモデルを用いた [9]. 元となる文章を分かち書きし, 学習済みモデルに入力すると 300 次元の分散表現が得られる. 数値データに変換した訓練データを入力データとして機械学習モデルの訓練を行う.

表 3　本研究におけるジャンル分類

科学	文学
言葉	日本史
世界史	地理
公民	芸術
芸能	音楽
生活	スポーツ

3.2　機械学習手法

機械学習には既存アルゴリズムとしてよく知られるランダムフォレスト,k-近傍法, ロジスティック回帰, ナイーブベイズ,SVM, 確率的勾配降下法 (以下:SGD) を用いた他にも,300 次元の入力層,1000 次元の中間層,12 次元の出力層からなる 3 層のニューラルネットワーク (以下:NN) も用いた. また, これらの手法に対してグリットサーチによるハイパーパラメータの最適化も行って評価した.

3.3　評価

精度の評価にはマクロ平均による正解率 (macro-Accuracy) を用いる. 多クラス分類で使う評価指標は, 二値分類で用いる混合行列を複数のクラスに対応させて算出できる.

たとえば (A, B, C) からなる多クラス分類を考えたとき,A を $Positive$, B と C を $Negative$ とした際の TP, TN, FP, FN をそれぞれ TP_A, TN_A, FP_A, FN_A とする. これらを用いて $accuary_A, recall_A, precision_A, F$ 値 $_A$ は式 (1)~(4) のように表すことができる.

$$accuary_A = \frac{TP_A + TN_A}{TP_A + TN_A + FP_A + FN_A} \tag{1}$$

$$recall_A = \frac{TP_A}{TP_A + FN_A} \tag{2}$$

$$precision_A = \frac{TP_A}{TP_A + FP_A} \tag{3}$$

$$F_A = \frac{2 \cdot recall_A \cdot precision_A}{recall_A + precision_A} \tag{4}$$

これらの値を各クラス毎に計算できる. これらの平均が次式のマクロ平均である.

$$accuary_M = \frac{accuary_A + accuary_B + accuary_C}{3} \tag{5}$$

Accuracy は不均衡データの評価には適さないことで知られるが, 本研究での分類問題は偏りが少ないことから適用する判断をした.

4　実験

4.1　実験結果

それぞれのモデルでグリッドサーチを行った. k 分割交差検証の分割数は k=5, 評価指標はマクロ平均 (macro-Accuarcy) とした.

表 5　機械学習モデルの正解率

機械学習モデル	正解率
ランダムフォレスト	0.696
k-近傍法	0.701
ロジスティック回帰	0.740
ナイーブベイズ	0.685
SVM	0.768
SGD	0.746
NN	0.720

4.2　考察

単一のモデルでは SVM が最高の正解率を得た. 正解率のみを求めるならば, アンサンブル学習を適用すれば精度の向上は見込める. ここで, 各モデルにおける 1 つのジャンルの正解率を表 6 にまとめた. 全体としてはスポーツの正解率が最も高く, 次いで科学, 文学, 地理が次点となった. 一方で, 音楽と芸術の正解率が 0.60 を下回る分類結果となっている. 特に音楽と芸術はジャンルとしても非常に近い関係性にあるといえる. たとえば,JPOP や洋楽は音楽に分類されるが, クラシックは芸術に分類されることが多い.

また, 実験中に生じた問題を紹介する. クイズの問題には 1 つのジャンルしか付与されていない. 付与されたラベルを「正解ラベル」, 解答が真に含まれるジャンルを「真のラベル」とする. データセット内には表 7 のような問題が含まれており, それらがノイズになっている可能性がある. そもそもジャンルの付与は作問者それぞれが個別で行うため, 同じ問題を作ってもジャンルが異なることがある. たとえば表 7 の 1 番の問題は, 前情報ではスポーツに関する情報, 後情報と解答は地理に関する情報である. このような問題が含まれることから, マルチラベル分類の適用が必要と感じた. 従って, 今後はマルチラベル分類手法を実装していきたい.

表 6　各機械学習モデルの F1 値

	音楽	日本史	世界史	生活	科学	芸術	スポーツ	言葉	文学	地理	芸能	公民
ランダムフォレスト	0.40	0.68	0.66	0.66	0.77	0.43	0.85	0.64	0.73	0.76	0.71	0.74
k-近傍法	0.61	0.61	0.64	0.66	0.73	0.54	0.86	0.64	0.74	0.74	0.72	0.67
ロジスティック回帰	0.60	0.77	0.72	0.71	0.77	0.62	0.90	0.66	0.76	0.77	0.75	0.69
ナイーブベイズ	0.61	0.61	0.64	0.66	0.73	0.54	0.86	0.64	0.74	0.74	0.72	0.67
SVM	0.65	0.75	0.74	0.69	0.79	0.62	0.90	0.70	0.80	0.78	0.79	0.77
SGD	0.55	0.75	0.72	0.69	0.74	061	0.89	0.59	0.76	0.75	0.76	0.68
NN	0.55	0.77	0.68	0.69	0.76	0.61	0.88	0.64	0.77	0.75	0.72	0.67
average	0.57	0.71	0.69	0.68	0.76	0.56	0.88	0.64	0.76	0.76	0.74	0.70

表 7　ジャンルに不備がある問題例

ID	問題	解答	正解ラベル	真のラベル
1	FIFA・国際サッカー連盟が本部を置く、スイスの都市はどこでしょう？	チューリッヒ	スポーツ	地理
2	ケインズの登場とともに成立した、財やサービスの全体集合を扱う経済学の一分野を、ミクロ経済学に対して何というでしょう？	ミクロ経済学	文学	公民
3	「無線 LAN」などというときの「LAN」とは、何という英語の略でしょう？	Local Area Network	科学	言葉

5　おわりに

先行研究で紹介したジャンル分類をその他の機械学習手法を試した. 結果としては先行研究でも使われた SVM が最高の正解率を記録した. 今後正解率の向上を目指すならばアンサンブル学習を適用すれば良いと思う. また, 実験よりマルチラベル分類が必要であることが分かったため, 今後はマルチラベル分類手法を調査して行きたいと思う.

謝辞

本研究で使用したデータセット用のクイズ問題は，abc/EQIDEN 実行委員会より研究目的での利用許可を頂きました. 記して感謝いたします.

参考文献

[1] 伊東栄典淀川翼. 機械学習手法を用いたクイズ問題のジャンル推定. 火の国情報シンポジウム論文集, Vol. 2020, pp. 1–4, 2020.

[2] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018. https://arxiv.org/abs/1806.03822.

[3] Mandar Joshi, Eunsol Choi, Daniel Weld, , and LukeZettlemoye. Triviaqa: A large scale distantly supervisedchallenge dataset for reading comprehension, 2017.

[4] Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and and-Jordan L. Boyd-Graber. Quizbowl: The case for incrementalquestion answering. Technical report, 2019.

[5] 鈴木正敏, 松田耕史, 岡崎直観, 乾健太郎. 解答可能性付き読解データセット. 言語処理学会第 24 回年次大会 (NLP2018), March 2018.

[6] 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 (NLP2020), March 2020.

[7] David Ferrucci, Eric Brown, Jennifer Chu-Carroll, JamesFa, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building watson: An overview of thedeepqa project. **AI Magazine**, pp. 59–79.

[8] 橋元佐知, 佐藤理史, 宮田玲, 小川浩平. 競技クイズ・パラレル問題の基本構造と文型. 言語処理学会第 27 回年次大会 (NLP2021), March 2021.

[9] 日本語 WIKIPEDIA で学習した DOC2VEC モデル, (2021-12 閲覧). https://yag-ays.github.io/project/pretrained_doc2vec_wikipedia/.