機械学習手法を用いたクイズ問題のジャンル推定

淀川 翼[1,a)]　伊東 栄典[2]

概要：問題に対する一意な回答から構成されるクイズは，人類が古来より楽しんできた知的娯楽である。クイズの形式は教育や学習における試験でも活用されている。クイズは知性を実現するものとも考えられており，人工知能の例としてクイズ回答 AI も作成されている。一方，クイズをスポーツ的な対戦および観戦として楽しむ，頭脳戦心理戦としてのクイズ大会も開催されている。クイズ大会で選手が勝ち残るため，選手はクイズの訓練を行う。クイズの訓練には，過去から現在までに作成され蓄積されたクイズ問題を利用できる。数多くのクイズから，適切な訓練問題を選出するには，クイズのジャンル分類や，難易度の数値化が必要である。本研究では，機械学習手法によるクイズのジャンル推定を行う。ジャンル推定手法および小規模実験結果について述べる。

キーワード：クイズ, 機械学習, ジャンル推定

Quiz genre estimation by machine learning

Tubasa Yodogawa[1,a)]　Eisuke Ito[2]

Abstract: A quiz consists of a question and unique answer. Human beings have enjoyed quiz as an intellectual entertainment since ancient times. The quiz format is also used in exams in teaching and learning. Quiz answering AI was created as an example of artificial intelligence. Recently, quiz tournaments are held, and guiz game is played as a sports game in the tournament. Players fight as a mental/psychological war, audience enjoy their games. Players train and learn quizzes to win the quiz tournament. For quiz training, player uses a lot of quizzes which are created from the past to the present. To select an appropriate training quiz, it is necessary to categorize the quiz genres and to quantify the difficulty. In this study, we propose an automatic genre estimation method using machine learning. This paper shows the method and some results of small-size experiment.

Keywords: Quiz, Machine learning, Genre estimation

1. はじめに

問題に対する一意な回答から構成されるクイズは，人類が古来より楽しんできた知的娯楽である。クイズの形式は教育や学習における試験でも活用されている。クイズは知性を実現するものとも考えられており，人工知能の例としてクイズ回答 AI も作成されている。

従来よりテレビでクイズ番組が放映されてきた。クイズは多くの人が気軽に参加でき，共感できるため，テレビとの親和性が高い近年では，クイズをスポーツ的な対戦および観戦として楽しむ，頭脳戦心理戦としてのクイズ大会も開催されている。日本の主に大学生を対象とした「全日本クイズリーグ（通称 AQL）」には，多くの大学生がクイズ大会に参加している。表 1 に，2017〜2019 年における AQL 全国のクイズ大会の参加団体数および参加者数を示す。

表 1　AQL 全国クイズリーグ大会参加者

大会	団体数	参加者数
プレ AQL (2017 年)	110	869
AQL2018	147	1,284
AQL2019	214	1,895

[1] 九州大学大学院ライブラリサイエンス
[2] 九州大学情報基盤研究開発センター
[a)] yodogawa.tsubasa.543@s.kyushu-u.ac.jp

クイズは娯楽だけでなく，脳の活性化にも役立つ。子供の教育現場における試験は，クイズ的な要素を持つ。クイズは子供の知的能力をアピールする場として使われることも有る。近年の日本では高齢者が増加している。クイズでは体力を使わず頭脳を使うため，高齢者でも気軽に参加できるレクレーションである。高齢者の脳の活性化にもなると考えられる。

クイズ大会で選手が勝ち残るため，選手はクイズの訓練を行う。クイズの訓練には，過去から現在までに作成され蓄積されたクイズ問題を利用できる。数多くのクイズから，適切な訓練問題を選出するには，クイズのジャンル分類や，難易度の数値化が必要である。子供や高齢者を含め，多様な人がクイズを楽しむには，その人の嗜好に合うクイズの選出（推薦）も必要である。

本研究では，機械学習手法によるクイズのジャンル推定を行う。クイズのデータセットとして，九州大学クイズ研究会が保持する約 3 万問の過去問を用いる。過去問にはジャンル情報が無いため，人手でジャンルを付与する。

本論文の構成を述べる。第 2 節ではクイズデータについて述べる。次に第 3 節で，クイズにジャンルを人手で付与するために構築した CGI システムについて説明する。第 4 節では，機械学習によるクイズのジャンル推定について述べる。最後に第 5 節でまとめと今後の課題を述べる。

2. 用いるクイズデータ

本研究で用いるクイズのデータは，九州大学クイズ研究会が保持する約 3 万問の過去問である。これらの多くはクイズデータを蓄積・公開するサイト [2] に有る問題を整形したものである。このサイトによると，クイズ大会 abc は「新世代による基本問題実力 No.1 決定戦」，クイズ大会 EQIDEN は「新世代の学生サークルによる、早押しクイズ日本一決定戦トーナメント」である。

表 2 にクイズの例を示す。

表 2 クイズの例

ID	問題	回答
1	アルファベットの「A」「B」「C」のうち、ローマ字で用いたときに唯一母音として扱われるものはどれでしょう？	A
2	アメリカのバラク・オバマ大統領と、日本の鳩山由紀夫首相に共通する、所属政党は何党でしょう？	民主党 [Democratic Party]
3	「相乗積・浸透圧・円周率」を表すときに共通して使われるギリシャ文字は何でしょう？	π

今回の分析では，分析対象が適度な範囲に絞られていること望ましい。クイズ大会に出題された問題は，クイズ問題の品質が均質である。主に大学生が対象であるため，クイズの難易度は大学生が理解できる度合いに調整されている。クイズの場合，必ず答えが 1 つに決まることが必要である。クイズ大会の過去問は，各大学にいるクイズ部員が多人数で検査しているため，クイズ問題の内容的品質も保たれている。

3. クイズへのジャンル付与 CGI

表 2 に示すように，クイズの問題と回答は含まれるものの，ジャンルは無い。分類問題で用いる機械学習手法では，分類器の訓練に用いる正解データが必要である。そこで，クイズに人手でジャンル（ジャンルタグ，単語）を付与することにした。

本論文の筆者だけでは，時間と労力の限界から 3 万問のクイズにジャンルを付与できない。また 1〜2 名のジャンル付与では偏る可能性もある。クイズ研究会員のように，クイズ分野に詳しい複数の人でジャンルを付与するほうが信頼できる正解データが得られると判断した。そこで，複数人によるジャンル付与システムを構築する。

図 1 に構築中のジャンル付与 CGI システムの概要を示す。

システムでは最初に利用者認証を行う。認証後，その利用者によりジャンル付与されたクイズの一覧を表示する。ボタンによりジャンル付与済みクイズの再編集，ジャンル未付与のクイズへのジャンル付与を選択する。「ランダムに 10 個選ぶ」を押すと，ジャンル未付与のクイズから，ランダムに 10 問を選択肢，ジャンル編集画面に遷移する。ジャンル編集画面では，クイズを選択すると，問題と回答の詳細が表示される。下にジャンル入力欄があるため，そこにジャンルとなる単語を入力する。

ジャンルとなる単語は複数個入力可能とする。基本的な文書のジャンル分けの場合，ジャンルは 1 つに限定することが多い。クイズや文書の場合，ジャンルを 1 つに絞ることが難しい。人手の作業を複数回行う労力を省くため，複数個の単語をジャンルに指定可能にしておく。

4. 機械学習によるクイズのジャンル分け

本研究の主要部分である機械学習によるクイズのジャンル分けを説明する。

4.1 Bag-of-Words によるクイズのベクトル化

機械学習で分類問題を解く場合，分類対象を数値のベクトルで表現する。文書のジャンル分けの場合，文書をベクトル化する必要がある。

最も基本的な文書のベクトル化手法は，Bag-of-Words を用いる方法であろう。図 2 にベクトル化の概要を示す。

日本語文書の場合，形態素解析により文を単語に分割する。形態素解析ツール Mecab では，辞書を指定することで新語にも対応可能である。単語に分割された後，不要語を削除する。日本語では「の」や「こと」のような補助語

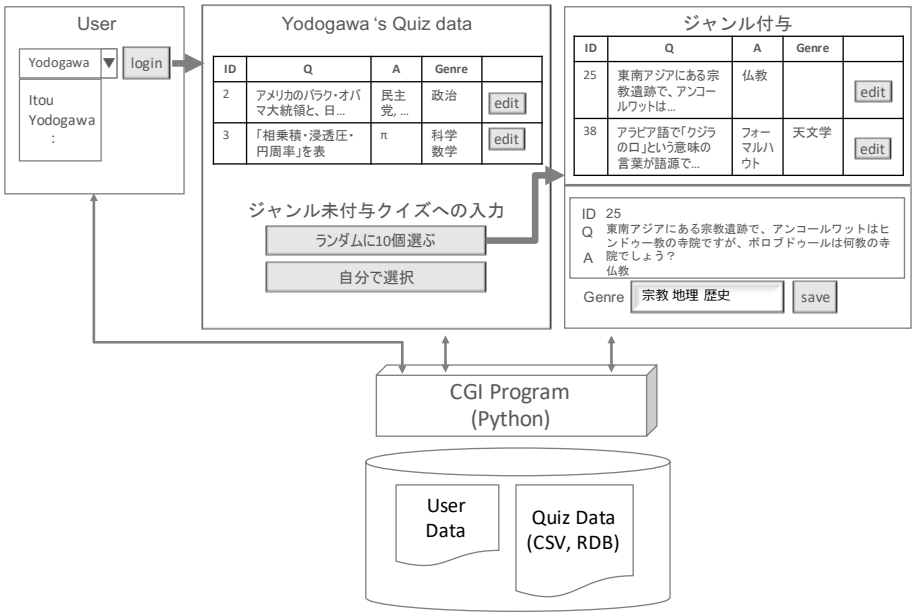

図 1　人手によるジャンル付与 CGI

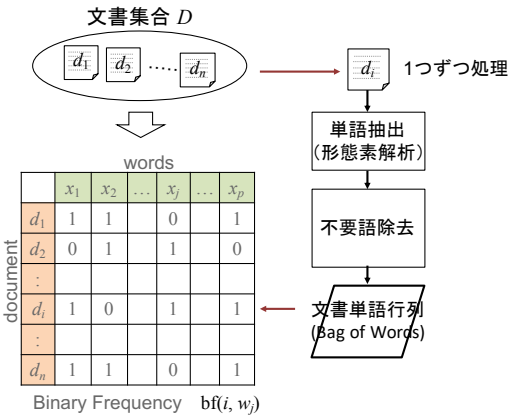

図 2　Bag-of-Words によるベクトル化

が多数出現する。これらの補助語は意味の解析では雑音となる。名詞のみに絞ることで，意味に基づく分析が出来る。

不要語削除の後，単語の出現頻度を数え，文書と単語の対応を表す文書単語を構築する。一般的な文書単語行列では，単語の出現頻度である TF (Term Frequency) を行列の要素値にすることが多い。文書 d_i に単語 x_j が出現した回数が行列の i, j の値になる。

本研究では行列の値として，単語が出現したか否かを表す BF (Binary Frequency) を用いる。BF を用いる理由は２つある。１つ目の理由はクイズ文の短さである。クイズ文は短いため，複数回出現する単語は少ない。そのため TF と BF の値に差が出ない。２つ目の理由は機械学習分類器に用いる際の精度である。筆者の経験上，TF よりも BF にした方が精度が高い場合が多い。そのため，今回も BF 値を用いて実験する。

4.2　単語の分散表現によるクイズのベクトル化

近年，Word2Vec などの単語の分散表現を用いた研究がなされている。Word2Vec は Tomas Mikolov らの開発した分散表現を生成する手法で，各単語を高次元のベクトルで表現する [3]。各単語を高次元ベクトルで表す分散表現では，単語のベクトルの加法・減法の結果が，単語の意味の加法・減法が成り立つ規則性が示されている。我々は単語の分散表現を用いた，クラスタリングの部分文書集合の内容推定手法を研究してきた [5]。そこで，単語の分散表現を用いたクイズ文書のベクトル化も検討する。

図 3 に fastText を用いた単語の分散表現獲得を示す。大規模な文書コーパスから抽出した文章を，形態素解析に分かち書きにする。分かち書きで単語の並びになった文章を fastText に入力し，単語の分散表現を得る。分散表現としてのベクトル長は 100〜300 を指定できる。図 3 では 300 として出力している。

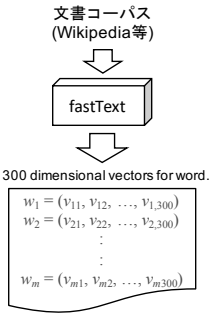

図 3　fastText による単語の分散表現獲得

次に文書のベクトル化手法を考える。最も簡単な方法として，文書に出現する単語のベクトル値の平均値を，その文書のベクトルとする方法が有る。図 4 に単語のベクトル

値の平均値を得る手順を示す。

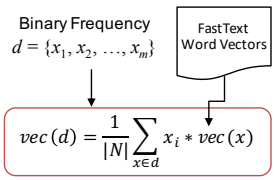

図 4　単語の分散表現によるベクトル化

4.3　分類器によるジャンル推定

文書を数値のベクトルで表現できれば，機械学習による分類器を構築できる。分類問題に使える機械学習分類器は，SVM (Support Vector Machine)，NN (Neural Network)，ランダムフォレスト等がある。図 5 に SVM による分類器作成を示す。

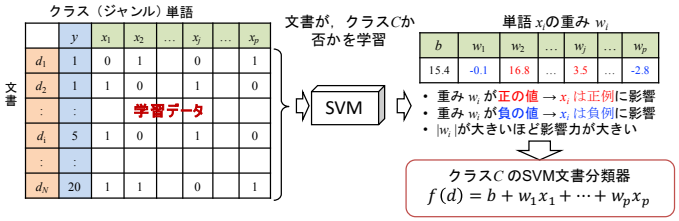

図 5　SVM によるジャンル推定

今回はクイズのジャンルを複数付与する。複数ジャンルを許容する場合，SVM および NN では問題無い。SVM は 2 値分類であるため，ジャンル数と同じ個数の文書ジャンル分類器を作る。1 つのクイズに 2 個のジャンルが付与されている場合，どちらのジャンルの分類器にも適合するように訓練されれば良い。

NN の場合，出力層のノード数をジャンル数と同数にすれば良い。例えば 2 個のジャンルを付与されたクイズ文書の場合，その 2 つのジャンルの出力層ノードの値が大きくなるようにすれば良い。

決定木やランダムフォレストでは，複数ジャンルの分類には適さないかもしれない。SVM, NN を含め複数の手法を試し，最も性能の良い分類器を採用する。

4.4　分類器の評価

用いるデータには約 3 万個のクイズ問題・回答が含まれる。すべてのクイズにジャンル付与が出来れば，3 万個のクイズで 5 分割交差検定，あるいは 10 分割交差検定で性能を評価する。性能は Precision, Recall, Accuracy, F1 score (F-Measure) で評価する。

5.　おわりに

本研究では，クイズ問題・回答の文章から，クイズのジャンル推定について述べた。九州大学クイズ研究会が保持する過去のクイズ問題 3 万個を用いて機械学習する。正解ジャンルのデータは無いため，人手で付与する。人手による作業を支援するための CGI プログラムを構築している。

クイズ文書のベクトル化手法として，BF による文書ベクトルと，単語の分散表現を用いたベクトル化を提案した。クイズ文書のベクトル化後，機械学習で SVM や NN によるジャンル分類器を作る。出来た分類器でジャンルが推定可能である。

今回は手法の提案だけで，実データにおる実験に至っていない。今後はデータの整備と，データによる実験を行う予定である。

参考文献

[1] AQL:全日本クイズリーグ, https://www.quizaql.com/ (accessed at 2020-02-25).
[2] ABC, http://abc-dive.com/ (accessed at 2020-02-25).
[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean: Distributed Representations of Words and Phrases and Their Compositionality, NIPS'13, vol.2, pp.3111–3119, 2013.
[4] Piotr Bojanowski,Edouard Grave, Armand Joulin, Tomas Mikolov: Enriching Word Vectors with Subword Information, arXiv preprint, arXiv:1607.04606, 2016.
[5] 淀川翼, 加登一成, 伊東栄典: 単語の分散表現を用いた文書クラスタのラベル推定, 人工知能学会 第 49 回セマンティックウェブとオントロジー研究会 SIG-SWO, vol.49, no.3, Nov. 2019.