ライブコンペティション：「AI王～クイズ AI 日本一決定戦～」

鈴木 潤^{†,††}・松田 耕史^{††,†}・鈴木 正敏^{†,†††}・加藤 拓真[†]・
宮脇 峻平[†]・西田 京介^{††††}

1 はじめに

本稿では、言語処理学会第 27 回年次大会に併設されたワークショップの一つ「AI 王 ～クイズ AI 日本一決定戦～」内で実施したクイズ AI のライブコンペティションについて、その概要と運営実態について紹介する。

2 コンペティション概要

2.1 ねらい

質問応答は、自然言語処理研究における主要な研究テーマの一つである。どのような質問にも的確に答えられるシステムの実現は、多くの自然言語処理研究者の夢であり、究極的なゴールの一つと言える。現在、ニューラルネット／深層学習による方法論の流行とともに、質問応答研究も世界的に大きな盛り上がりを見せている。事実、人工知能関連の難関国際会議では、毎年多くの質問応答に関する研究成果が報告されている。しかし、最先端技術にキャッチアップしながら継続的に質問応答の最先端研究に従事している日本の研究者・研究機関は非常に少なく、世界的な流行との乖離が見られる。深層学習時代以降の日本での質問応答研究の促進のため、「日本語によるクイズ」を題材とし、最新の質問応答技術に触れながら技術や知識を習得する場を提供することを目的として本コンペティションを開催することを計画した。特に日本では、昨今世間的にクイズ番組が一定の流行をみせており、若手研究者が興味を持ちやすい題材と考えられる点を勘案してこの題材を選んだ。また、日本語の質問応答データを提供することで、日本語に対する研究を促進するという側面も考慮している。

2.2 ルール

本コンペティションの基本的な理念は、これまでに開催された機械学習系の共有タスク (shared

[†] 東北大学

^{††} 理化学研究所

^{†††} 現所属：株式会社 Studio Ousia

^{††††} 日本電信電話株式会社 NTT メディアインテリジェンス研究所

task) でしばしば採用されていた, 「配布データのみでシステムを構築」という設定ではなく, 「あらゆるデータを使ってクイズの正解率を向上させること」とした. ただし, 誰でも取得可能な無償公開されているデータのみを対象とするという制限を設けた. これは, 本コンペティションの目的が主に若手研究者を対象とした研究促進であり, 一部の参加者 (特にデータを大量に保持している組織) しか使えないデータを許すと, 公平感や再現性の担保が難しく, 本コンペティションの主旨に反するという考えに基づいた措置である¹. 計算機リソースに関しても同様に, 基本的に単一の計算機内のリソースで動作するシステムとした².

2.3 評価

質問応答のコンペティションをする上で避けて通れない課題の一つとして, 答えの表記揺れや別解への対応がある. つまり, システムが出力する解答の正誤を判定する際に, 正解文字列との完全一致度で判断すると, 本来は正解だが異表記となるものを誤答と判定してしまう可能性がある. この課題に正面から対応するためには, 人間による人手チェックが必要となる. しかし, 正誤判定にかかる時間的・金銭的成本は無視できないほど大きく, 特に, 今回実施したライブコンペティションの場合は, 解答提出から表彰までの時間は短く, その間に参加システムの全解答を人手チェックすることは難しい. そこで, 本コンペティションでは, 解答の選択肢を設け, その選択肢から正解を選択する問題設定とした. このような選択肢付きの択一問題とすることで, 正解を唯一に決めることが可能となり, 正誤判定における曖昧性や判定誤りの発生をなくすことができる³. また, 単純な選択肢の一致度だけで正解率が計算できるため, 時間的・金銭的成本なく評価でき, ライブコンペティションに適した設定と言える.

2.4 評価用データセット

表 1 に, 評価用データセットに含まれるクイズ問題の例を示す. 1 つのクイズ問題は, 質問・選択肢・正解の 3 つ組からなる. 質問および正解は, 本コンペティション用にクイズ作家に新

質問: 北海道の中心に位置することから「北海道のへそ」を名乗る、ラベンダーで有名な都市はどこ？
選択肢: 1. 名寄市, 2. 紋別市, 3. 富良野市, 4. 当別町, ..., 18. 南富良野町, 19. 赤平市, 20. 士別市
正解: 3. 富良野市

表 1 クイズ問題の例

¹ データを無償公開する場合は, 独自に作成したデータであっても利用可能とした.

² どこまでを単一計算機と判断するかは明確な規定は設けず常識的な範囲で考える, としている. また, 「単一の計算機内のリソース」なので, 実行時のインターネット検索は利用禁止とした.

³ 評価データの選択肢に関しては, 事前に人手により全数チェックを行い, 正解以外の選択肢が別解となっていないことを確認した.

たに発注し作成した。基本的には、クイズ番組やクイズイベントで作成する方法と同じ手順での問題作成を依頼した。ただし、前述の通り解答の表記揺れなどを防ぐために、クイズ問題の正解は必ず日本語 Wikipedia の記事名になること、という条件を設けて作問を依頼した。

評価データ作成手順 まず、日本語 Wikipedia の記事名一覧を取得し、その中からリダイレクト元のページなどを取り除いた約 92 万の記事名のリストを取得した。次に、取得した記事名リストをクイズ作家に渡し、前述の通りリスト中の記事名が正解になるクイズ問題の作成を依頼した。クイズ作家からクイズ問題が納入された後、それぞれの質問・正解ペアに対して、正解を含む 20 個の選択肢を最初に取得した約 92 万の記事名リストから選出し付与した。したがって、全ての選択肢は必ず日本語 Wikipedia の記事名であることが保証されている。

選択肢付与手順 初めに「日本語 Wikipedia エンティティベクトル」(Suzuki et al. 2018)⁴を用いて、正解との類似度の高い 5 件の Wikipedia 記事ベクトルのランダムな重み付き和であるベクトルを求める。次に、求めたベクトルとの類似度が上位である記事のうち、「拡張固有表現 ver.8」(関根 他 2020) のクラスが同じものを優先して 19 個選び、正解と合わせて 20 個の選択肢とした。このようなランダム性を導入した 2 段階の類似度検索によって、解答時に選択肢のベクトルの重心を計算することで正解を導くような短絡的な解法を防ぐ工夫をしている。また、正解と固有表現クラスが同じ選択肢をなるべく多くすることで、質問で問われている固有表現クラスの解析だけでは質問に解答できないようにした⁵。なお、本アルゴリズムは、コンペティションが終了した後に参加者に開示した。

3 幅広い参加者を集めるために実施した施策・工夫

訓練・開発用データなどの配布 筆者らが以前に公開した日本語による質問応答のデータセット JAQKET (鈴木 他 2020) の訓練用データと開発用データを参加者に配布し、モデルの訓練と性能検証に利用できるようにした⁶。また、上記約 92 万の記事名のリストとその各記事名に対応する Wikipedia 記事から本文を抽出したものも合わせて公開した。これらを用いることで、参加者はデータを独自に集めなくても最低限のクイズ AI の学習を行うことができる。

ベースラインシステムの提供 東北大学研究室が公開した日本語用 BERT モデル⁷をベースにしたベースラインシステムを用意し、GitHub 上に公開した⁸。これは GitHub から clone してシェルスクリプトを動作させるだけで、必要なデータのダウンロード、モデルの学習、(後述す

⁴ <https://github.com/singleton/WikiEntVec>

⁵ ただし、正解と同じ固有表現クラスのエンティティが必ずしも多数存在するとは限らないため、表 1 の選択肢にある「当別町」などのように、正解との意味的類似度は高いが固有表現クラスが異なるような選択肢も一部存在する。

⁶ JAQKET の訓練データは、既存のクイズ問題に対して 2.4 節に記載の方法で選択肢の付与を行ったものである。

⁷ <https://github.com/cl-tohoku/bert-japanese>

⁸ https://github.com/cl-tohoku/JAQKET_baseline

図 1 リーダーボードへの投稿推移.

図 2 ライブコンペティションにおける各チームの正解数.

るリーダーボードに投稿可能な) 評価データの質問に対する解答の予測が自動的に行われ, 全くコードを書かなくてもベースラインの再現が可能となっている.

リーダーボードの運用 参加者に継続的に取り組んでいただけるよう, コンペティション初期の段階からリーダーボードを運用した. このリーダーボードには 22 チームから 400 回を超える投稿があり, 参加者のモチベーション維持に貢献したと思われる. また, リーダーボード自体の活性化のために, ライブコンペティション当日は評価データに対する表彰とは別に, リーダーボードの上位チームも顕彰した. 図 1 はリーダーボードに投稿された結果の正解率を表している. 個別の投稿をマーカーで表し, その時点での最高到達正解率を実線で表している. 継続的に投稿され, ライブコンペティション直前まで最高性能が向上し続けていることが見て取れる.

Slack/Twitter の活用 Slack ワークスペースをアナウンスや参加者同士の情報交換の場として活用し, 合計 68 名が利用した. Twitter では, 「AI 王〜クイズ AI 日本一決定戦〜」(@AioJaquet) というアカウントを開設して広報活動を行った. 開催案内や結果報告のツイートが 30,000 インプレッションを超えるなど一定の効果を得た.

参加費補助制度 本コンペティションは年次大会併設ワークショップとして実施したため, 参加者は年次大会本会議への参加登録および参加費の支払いが必須である. 学生においては, 研究室などからの補助が受けられず, 参加費を理由に参加を断念してしまう可能性が想定された. そこで, ライブコンペティションと参加システム報告の双方に発表参加することを条件に, 年次大会の事前登録参加費分相当を補助する制度を設けた⁹. 実際に, 3 名の学生参加者から申請があり, 参加費補助を行った.

4 ワークショップ (ライブコンペティション) 当日の運営

概要 最終的に, ライブコンペティションには 11 チームから参加申し込みを受けた. そのう

⁹ 補助金の原資は運営委員のポケットマネーである.

ち、9 チームからワークショップ内で参加システム概要の口頭発表があった。また、当日は3 件の招待講演を企画した。有益な招待講演の効果もあり、また、年次大会のワークショップとして開催することでライブコンペティションに参加しない聴衆も数多く参加していただくことができ、ピーク時で140 名ほど、ワークショップ全体を通して常時100 名以上の参加者を確認することができた。図2 に、ライブコンペティションの最終結果を示す。上位4 チームは正解率が9 割を超えるなど、非常に接戦であった。上位チームは、共通してBERT をベースとした上で、主な工夫として(1) 情報検索に基づく解答に必要な情報の正確な絞り込み、(2) データ拡張による問題の網羅率向上、(3) クイズの特徴を踏まえた質問文の修正、などを適用して正解率の向上を図っていた。

運営状況 COVID-19 感染拡大の影響で年次大会自体が、参加者にとってはオンライン開催となり、当然、本コンペティションもオンラインでのリモート開催・運営となった。さらに、運営の本拠地となった東北大学のある仙台でも緊急事態宣言が発令され、運営委員も各自別々の場所でリモート運営をすることとなった。そのため、運営委員それぞれがワークショップのZoom ミーティングとは別に、運営用のZoom ミーティングを繋ぎ（ハウリングがひどく同時にやりとりをするのは難しかったので、ベストプラクティスとは言い難い）、Slack も駆使して情報をやり取りしながら運営を行った。

評価手順 まずワークショップの冒頭に評価用のクイズ問題1,000 問を参加者に配布した。次に、ワークショップのプログラムとしては、招待講演2 件および昼休みの時間（合計3 時間）の裏で、各参加者に提出する解答を準備していただき、準備ができたところで、Slack を経由して解答の提出を頂いた。その後、招待講演および参加者による参加システム報告の時間（合計3 時間）に各参加者の解答評価および表彰式の準備を行った。

所感 急遽決まったリモート運営に対する不安はあったが、運営委員の努力と参加者の好意的な協力のもと、大きな問題はなく無事にライブコンペティションを終えることができた。第一回目の開催なので、企画当初はコンペティション参加チーム数6 チーム、ワークショップ参加者数50 名を目標としていたが、結果的には上記の通り予想を大幅に上回る参加チーム／参加者にお集まりいただくことができた。コンペティション終了後は運営と参加者の一部による簡単なオンライン意見交換会／打ち上げが自発的に実施されるなど、研究ネットワークの形成にも貢献できた。当初掲げた質問応答研究の促進という観点は十分に満たせたと考えている。

5 おわりに

日本における質問応答研究の促進と活性化を目的として、言語処理学会第27 回年次大会に併設のワークショップ内で試行的に開催したライブコンペティションの報告を行った。本コンペティションは今年度以降も引き続き行っていきたいと考えている。次回はより現実的なクイズ

問題の設定として、選択肢を与えない、真の意味でのオープンドメイン質問応答での設定を検討中である。本ワークショップの想定ターゲット層としていた、自然言語処理研究を初学者の学生や若手研究者へどれほどリーチできたかは未知数ではあるが、データとベースラインを公開していることで、今後も継続的に質問応答研究の導入として利用可能である。また、引き続き参加者も聴衆も楽しめるコンペティションを企画していきたいと考えている。

謝 辞

AI 王プロジェクトは、理化学研究所 革新知能統合研究センター (<https://aip.riken.jp>) の支援を受けて実施しています。また、本コンペティションの評価用データのクイズ問題は、株式会社キュービックおよびクイズ法人カプリティオへ依頼して作成しました。記して感謝いたします。

参考文献

- 関根聡, 安藤まや, 小林暁雄, 隅田飛鳥 (2020). 拡張固有表現定義の更新と日本語 Wikipedia 分類データ 2019. 言語処理学会第 26 回年次大会, pp. 1221–1224. [S. Sekine et al. (2020). Update of Extended Named Entity Definition and Categorization Data on Japanese Wikipedia 2019 version. Proceedings of the 26th Annual Meeting for the Association for Natural Language Processing, pp. 1221–1224.].
- 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也 (2020). 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. [M. Suzuki et al. (2020). JAQKET: Constructing a Japanese Question Answering Dataset from Trivia Questions. Proceedings of the 26th Annual Meeting for the Association for Natural Language Processing, pp. 237–240.].
- Suzuki, M., Matsuda, K., Sekine, S., Okazaki, N., and Inui, K. (2018). “A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles.” *IEICE Transactions on Information and Systems*, **E101.D** (1), pp. 73–81.

略歴

- 鈴木 潤: 東北大学データ駆動科学・AI 教育研究センター教授。理化学研究所革新知能統合研究センター自然言語理解チーム客員研究員。
- 松田 耕史: 理化学研究所革新知能統合研究センター自然言語理解チームテクニカルスタッフ。東北大学大学院情報科学研究科博士後期課程在学中。
- 鈴木 正敏: 株式会社 Studio Ousia システムエンジニア。2021 年 3 月東北大学

大学院情報科学研究科博士後期課程修了.

加藤 拓真：東北大学大学院情報科学研究科博士前期課程 2 年.

宮脇 峻平：東北大学大学院情報科学研究科博士前期課程 2 年.

西田 京介：日本電信電話株式会社 NTT メディアインテリジェンス研究所 特別研究員.