QASports: A Question Answering Dataset about Sports

Pedro Calciolari Jardim[1], **Leonardo Mauro Pereira Moraes**[1,2], **Cristina Dutra Aguiar**[1]

[1]Department of Computer Science, University of Sao Paulo, Sao Carlos, Brazil

[2]Data & AI Center of Excellence, Amaris Consulting, Vernier, Geneva, Switzerland

{pedrocjardim, leonardo.mauro}@usp.br, cdac@icmc.usp.br

Abstract. *Sport is one of the most popular and revenue-generating forms of entertainment. Therefore, analyzing data related to this domain introduces several opportunities for Question Answering (QA) systems, such as supporting tactical decision-making. But, to develop and evaluate QA systems, researchers and developers need datasets that contain questions and their corresponding answers. In this paper, we focus on this issue. We propose QASports, the first large sports question answering dataset for extractive answer questions. QASports contains more than 1.5 million triples of questions, answers, and context about three popular sports: soccer, American football, and basketball. We describe the QASports processes of data collection and questions and answers generation. We also describe the characteristics of the QASports data. Furthermore, we analyze the sources used to obtain raw data and investigate the usability of QASports by issuing "wh-queries". Moreover, we describe scenarios for using QASports, highlighting its importance for training and evaluating QA systems.*

1. Introduction

Sport is a topic of great interest and is constantly growing thanks to its popularity and revenue. According to a recent report[1], the sports industry became a global driver of the economy; its estimated size is $1.3 trillion, and its audience is over 1 billion people. People enjoy sports for different reasons, such as supporting their favorite teams by attending matches, watching matches on streaming services, betting online or offline, practicing sports out of passion, or playing video games for entertainment.

The sports domain introduces several significant computational opportunities for database systems and artificial intelligence [Beal et al. 2019]. Examples include match outcome prediction, tactical decision-making, player investments, and injury prediction. However, a large amount of sports data is available, requiring specialized systems to extract information from these data efficiently. Question Answering systems can provide a suitable solution for this challenge.

A Question Answering (QA) system stores and processes several different documents formats, such as web-based files, to extract information through questions[2] written in natural language [Karpukhin et al. 2020, Moraes et al. 2023]. It also provides a unified and accurate way to query textual documents by employing specialized algorithms usually composed of two steps. The first step, Document Retriever, receives questions

[1]Sports Industry Statistic and Market Size Overview, Business and Industry Statistics. May 5, 2023.

[2]In this paper, we use the terms *query* and *question* interchangeably.

in natural language from the users and searches for relevant documents that can provide suitable textual data to these questions. The second step, Document Reader, produces summarized answers from the retrieved documents.

In addition to support questions in natural language, QA systems also understand context, subject, and question intention [Mishra and Jain 2016, Karpukhin et al. 2020]. Thus, they differ from search engines, which provide a list of relevant documents based on factors like popularity, keywords, and frequency of access and require users to manually examine each document to find specific information. As a result, QA systems tend to be more efficient than search engines [Athira et al. 2013, Mishra and Jain 2016], introducing advantages for many applications. For instance, we can use a QA system to create a cutting-edge search engine specialized in analyzing legal documents, or an intuitive FAQ chatbot that supports users queries related to products and services.

To develop and evaluate QA systems, researchers and developers need datasets that contain questions and their corresponding answers. These datasets typically consist of pairs of questions and answers or triples of questions, answers, and context. There are different types of QA datasets, depending on their characteristics. Some datasets are general-purpose, containing questions and answers on a wide range of topics, while others are more specialized, focusing on specific domains. Furthermore, the QA datasets can contain multiple choice questions or extractive answer questions. In the former, the dataset stores a set of alternatives for each question. In the latter, the dataset stores one or more sentences for each question and its respective answer.

According to our discussions in Section 2, there are some QA datasets publicly available [Richardson et al. 2013, Rajpurkar et al. 2016, Nguyen et al. 2016, Hill et al. 2016, Lai et al. 2017, Kwiatkowski et al. 2019, Liu et al. 2020]. These datasets use many sources for their documents, including Wikipedia and books, covering a great variety of topics. Nevertheless, none of those datasets focuses on the sports domain with extractive answer questions, thus making it difficult to assist the sports decision-making in pure textual data analysis. In fact, specialized QA datasets are important for improving information retrieval quality in specific domains. Furthermore, these datasets only store a small volume of sports data, imposing limitations when training and evaluating QA systems. For instance, these systems can not learn about particular characteristics of the sports domain, such as terminology and context comprehension.

To the best of our knowledge, we present the *first* large sports question answering dataset for extractive answer questions, named QASports. QASports contains real data of players, teams, and matches from the following sports: soccer, basketball, and American football. It has over 1.5 million questions and answers about 54 thousand cleaned and organized documents from Wikipedia-like sources. Its final size is about 1.97 GB, allowing QA systems to learn about particular characteristics of the sports domain. Furthermore, it can be used to train and evaluate both the Document Reader and the Document Retriever of QA systems. QASports is publicly available for download (see Section 6).

This paper is organized as follows. Section 2 reviews related work, Section 3 introduces the proposed QASports dataset, Section 4 provides analyzes considering different aspects of QASports, Section 5 highlights different scenarios for using QASports, and Section 6 concludes the paper.

2. Related Work

Question answering datasets come in various forms, including multiple choice questions and extractive answer questions datasets. The most common type of QA dataset contains questions that can be answered with one or more sentences, or spans, of the document being interrogated, called extractive answer questions datasets. The answer is extracted from the context of a given question. The extraction tends to rely on matching the entity and the type of information the question requires with a phrase from the document. In some cases, a question may not have an answer. We classify SQuAD, AdversarialQA, Natural Questions, MS Marco, and Children's Book Test as this type of QA dataset.

SQuAD (Stanford Question Answering Dataset) [Rajpurkar et al. 2016] is a dataset composed of 536 Wikipedia articles and 107,785 context-question-answer triples, representing 35.1 MB of textual data. Its data refers to several domains, such as pharmacy, medicine, databases, software testing, TV series, and geology. The content of SQuaD was generated as follows. Crowdworkers were given a paragraph and tasked with asking and answering questions. They were discouraged from asking questions that were too similar to the information displayed in the context. Thus, the generation of SQuaD required a substantial manual effort. Meanwhile, AdversarialQA [Bartolo et al. 2020], a QA dataset in which humans have created adverse and complex questions, so the models cannot answer these questions easily. It is similar to SQuaD, although it is smaller. It counts with 72k questions in 36.1 MB of data.

Natural Questions [Kwiatkowski et al. 2019] and MS Marco [Nguyen et al. 2016] contain questions obtained from queries issued against search engines. Natural Questions uses Google and presents the question and the most relevant Wikipedia pages related to the question. MS Marco uses Bing and includes on average ten relevant documents from any website. The content of these two datasets cover several domains. MS Marco counts 102,023 questions in 169 MB of data. Meanwhile, Natural Questions counts 315k questions in 45.07 GB. However, Natural Questions present the entire content of the HTML page as the context of the questions, rather than just the phrase / sentences like the other datasets, resulting in a larger than normal dataset.

Children's Book Test [Hill et al. 2016] includes fairy tales that utilize simple narrative frameworks that are designed to aid children in comprehending the stories. As mentioned previously, by using fictional stories, the questions can only be answered with information from the given context, and not much real world knowledge is required. Children's Book Test counts 687,451 questions in 603 MB of data.

Similar to the datasets described in this section, QASports is also an extractive answer questions dataset. But our proposed dataset stores over 1.5 million of questions about 284k documents in 1.97 GB of preprocessed data. On the other hand, SQuAD stores approximately 350 sports documents, corresponding to 1.52% of its total volume. Regarding to MS Marco, it stores about 400 sports documents, corresponding to 0.01% of its total volume. In summary, the data available is insufficient to effectively train specialized QA models in the sports domain. By focusing on this topic and extracting from specialized sports websites, our dataset has the potential to enhance the vocabulary and terminology in sports use cases. In comparison, our dataset contains approximately 20 times as many questions as the traditional and popular SQuaD dataset has in total.

2.1. Other sports datasets

Sports datasets are a valuable resource that offer coaches and researchers a wealth of information and insights to enhance their analysis. There are other types of sports datasets containing a wide range of data points, including player statistics, team performance, game outcomes, and historical records. Studying sports datasets helps researchers rank players or teams, and understand the factors that lead to success or failure in game matches. Overall, sports datasets offer a comprehensive view of the game and are essential for anyone seeking to explore and understand the intricacies of sports.

Player monitoring has become a common task in many sports. Soccer2014DS [Ribeiro et al. 2017] covers player events from the 2014 Soccer World Cup, specifically focusing on player monitoring tasks. This dataset is composed of the raw extracted data collected by a web crawler and by derived streams with new calculated attributes. It can be useful for diverse artificial intelligence areas, such as data mining, sports analytics, and continuous preference queries.

Predicting match results is another important task in sports domain. To construct innovative predictive models for soccer match results, it is imperative to have access to comprehensive databases that go beyond mere statistical information. SoccerNews2018 [Alvares and Ribeiro 2019] proposes a database containing statistical data and news about the 2018 Brazilian Soccer Championship teams. This dataset contains diverse applicability beyond match prediction, such as pattern mining, outlier detection and sentiment analysis. However, it is not suitable for use in a question answering task.

Using multiple choice questions datasets like LiveQA, models learn to make guesses about a set of alternatives to a given question. LiveQA [Liu et al. 2020] collects play-by-play NBA broadcasts from the Chinese website Hupu, where all match points (or mistakes) are annotated individually. Furthermore, users participate in quizzes during the match to predict the team that will win, how many points a given player will score, and how many times a particular event, like missing a free throw, will happen in the rest of the match. As a result, the answer may change depending on the moment the user receives the question. Also, to calculate the total number of scores for a given player, the model must use the match broadcast to track time-sensitive data and add up all the moments a player scores. LiveQA counts 1,670 question in 115 MB of data.

Different from the datasets described in this section, QASports is an extractive answer questions dataset. For instance, when dealing with multiple choice datasets, QA models usually focus on picking the right option for a question. Although these models have many applications, they are not aimed at comprehending a text. Our dataset provides support for decision-making based on analyzing textual data. In addition, our dataset comprises Gigabytes compared to Megabytes of the related datasets.

3. The Proposed QASports Dataset

In this section, we present QASports, the *first* large sports question answering dataset for extractive answer questions. This dataset is composed of three pieces: (i) a collection of 51,874 cleaned pages stored in 270 MB of JSON files; (ii) context files in 3 CSV files, one for each sport, counting 284k lines and 209 MB of data; and (iii) generated questions and answers resulting in 1.5 million context-question-answer triples in 1.97 GB

of data. Section 3.1 describes the data collection and Section 3.2 details the question and answering generation. Section 3.3 describes the data structure and organization.

3.1. Data Collection

Wikis are websites created and maintained by its own audience. Although storing user-provided content can present issues regarding data accuracy, these websites have the advantage of allowing for a communal effort to keep and evolve knowledge on specific topics, reaching massive volume proportions. For instance, some wikis have 100 thousand pages [Mittell 2009]. Another advantage is that these wikis contain data related to several domains, such as games, TV series, and sports.

We extracted textual data from sports wikis available on the Fandom wiki[3] hosting service. The data refer to three of the most popular sports in the world: soccer[4], American football[5], and basketball[6]. On these wiki websites, each page has a header, a footer, and a sidebar with internal and external links. Inside the page content, there is a specific HTML structure that were used for reference. To extract useful texts, we removed all external links and Ads. The wikis were created using the same hosting service, resulting in a similar structure. Therefore, collecting and cleansing all the pages followed a similar process. Figure 1 depicts the employed data collection stepts, described as follows.

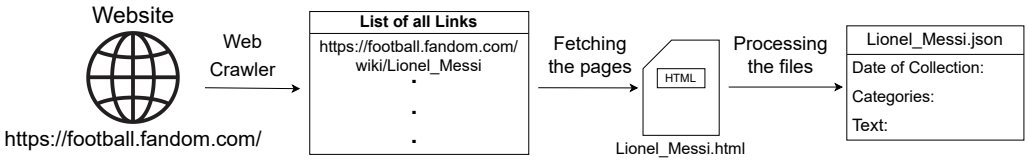

Figure 1. Data collection steps.

The first step is *list all link pages*. The soccer, American football, and basketball wikis have about 22k, 26k and 5k pages displayed on their header website, respectively. This is the total amount of current pages in each wiki. We collected all links available on the website by accessing the "All Pages" section, which lists every page on the wiki. It contains about 360 links listed in alphabetical order. We extracted all the links in the first page and repeated the process until the last sub-page.

Table 1 shows that we collected 29k, 27k, and 6k links for the soccer, American football, and basketball wikis, respectively. For each wiki, the number of collected links exceeds the number of displayed pages. We analyzed the pages and found duplicate links and pages without relevant textual content. An example of duplicate links is the case of Real Madrid's B team, also recognized as Real Madrid Castilla. Because of this duplicity, multiple links refer to the same page.

The next step refers to *fetching the raw HTML files*. To this end, we used a Python library to get the HTML (HyperText Markup Language) file from a wiki page. We kept these raw files for data preservation based on the following explanation. Raw files provide

[3]Fandom wiki - fandom.com
[4]Soccer wiki - football.fandom.com
[5]American football wiki - americanfootball.fandom.com
[6]Basketball wiki - basketball.fandom.com

Table 1. Gathering all links pages.

Sport	Displayed pages	Collected links	Difference
Soccer	22,518	29,435	+6,917
Football	26,602	27,679	+1,077
Basketball	5,310	6,652	+1,342
Total	54,430	63,766	+9,336

an original record of the processed data. Keeping raw files allows data preservation in its original form, which can be helpful for future reference, analysis, or verification.

The last step is *processing textual data from the HTML files*. We used Python's Beautiful Soup library to load the HTML and construct the DOM tree. From the DOM tree, we selected the valuable elements of the page by id or class name. We ignored irrelevant elements such as ads, figures, and sections with useless information. Most pages contained a small table with an overview of the player or the team, detailing specific information such as age, height, awards, and stats. We separated such information from the text and stored it in the JSON file under the key "infobox".

After cleaning and extracting the text from each page, we stored the resulting textual data in a JSON file. We also stored in this file related metadata such as page title, URL, collection date, and page categories. Figure 2 depicts an example of JSON output collected from a sport wiki page.


```
{
  "url": "https://football.fandom.com/wiki/Messi",
  "title": "Lionel Messi | Football Wiki | Fandom",
  "categories": [
    "1987 births", "Players", "Forwards",
    "Argentina international players",
    ...
  ],
  "date": "19/03/2023 15:43:37",
  "infobox": "Lionel Messi Personal information Full name Lionel Andres Messi Cuccittini
      Date of birth 24 June 1987 (1987-06-24) (age 35) Place of birth Rosario,
     Argentina Height 5 ft 7 in (1.70 m)...",
  "text": "General Image gallery Lionel Andres Messi (born 24 June 1987), also known as
     Leo Messi, is an Argentine professional footballer who captains the Argentina
     national team and plays for Ligue 1 club Paris Saint-Germain. [...]"
}
```


Figure 2. Output JSON of a sports wiki page.

3.2. Questions and Answers Generation

The first step to generate the questions and answers is to create the contexts, described as follows. Context in question answering tasks is the information or text that is given with a question. It can be a paragraph, a document, or even multiple documents. We divided the documents into smaller contexts by splitting them into sentences and adding them up to a minimal length of 256 characters. Thus, we maintained the connection between small sentences and minimized the chances the context does not provide the information independently. We generated over 285k contexts about the three sports: 49k, 88k, and 148k contexts for basket, soccer, and football, respectively.

We used Haystack[7] to generate questions and answers from a given context. The generation involved two steps: (1) We employed the T5 model (`t5-base`) [Raffel et al. 2019], which uses a neural network, to generate questions. It returned multiple questions for the context. (2) We then employed RoBERTa (`roberta-base`) [Liu et al. 2019], another neural network model, to answer the generated questions, including the possibility of a no-answer. This generation process produced a series of triple context-question-answer as the resulting output.

The use of the T5 and RoBERTa language models was motivated by the fact that, in 16 out of 18 categories analyzed, [Pan et al. 2023] demonstrated that large language models are a superior alternative to crowdworkers for data labeling. This results in cost savings of $500,000 dollars and 20,000 hours of work.

3.3. Data Description

The data description refers to details about the QASports structure and organization, such as folders and files. It can help ensure the data organization and prevent misunderstandings in the analysis and usability of the dataset.

First, we describe details about the *JSON files* gathered from the HTML pages. These documents capture specific data elements from the HTML pages, such as text content, metadata, structured information like tables and lists, and semi-structure data. We extracted the JSON documents from each of 51,874 wiki HTML pages and created documents according to the output shown in Figure 2. We stored the documents in the `Crawler/` folder. Each document contains the following attributes: (i) `url`: real URL; (ii) `title`: web page title; (iii) `date`: collection date; (iv) `categories`: categories of the page; (v) `infobox`: information about the players, teams and clubs; and (vi) `text`: textual data from the web page.

Table 2 depicts the number of documents we collected for each sport. The 51,874 wiki pages we managed correspond to about 11 GB of HTML pages. After the data and clean preprocessing, the JSON documents count about 270 MB. The collection, extraction and storage process took about 40 hours using a computer with 16GB RAM, 500 GB HD, Intel i5 11th Gen processor with 6 cores.

Table 2. JSON documents size

Sport	Documents	HTML (Disk Size)	JSON (Disk Size)
Football	26,549	4.9 GB	130 MB
Soccer	19,954	5.2 GB	112 MB
Basketball	5,371	0.9 GB	28 MB
Total	51,874	≈11 GB	≈270 MB

We now move our discussion to detail the creation of the *contexts*. A context is the knowledge source from which the questions are formulated. It also supports deriving the answers. We considered a specific list of contexts for each sport, separated as follows: `Contexts/Soccer.csv`, `Contexts/Football.csv`, and `Contexts/Basketball.csv` for soccer, football, and basketball, respectively. We divided the pages into smaller parts containing the attributes `url`, `title`, `date`, and

[7]Haystack - haystack.deepset.ai

`categories`. The process generated 88k, 148k, and 48k contexts for soccer, football, and basketball, respectively.

In the final stage of the dataset pipeline, we generated an output consisting of the following three large CSV files: `Question-answering/Soccer.csv`, `Question-answering`, and `Question-answering/Basketball.csv` for soccer, football, and basketball, respectively. As depicted in Figure 3, each CSV file contains a set of triples consisting of the context, the question, and the corresponding answer. Additionally, the files include related metadata like `qa_id` (i.e., question id), `context_id`, `context_title`, `context_categories`, and `url`. This generation took about 36 days using the same computational environment previously described. Together, these files store more than 1.5 million records in 1.97 GB of data.

```
1 qa_id, context_id, context, question, answer, ..., url
2 "25550...", "18109...", "Regional semifinals (Sweet Sixteen) Xavier, the third seed in
      the West, defeated seventh seed West Virginia,...", "How many points did B.J.
      Raymond score in the bonus round?", {'text': 'eight', 'offset': [73, 78]}, "https://
      basketball.fandom.com/wiki/..."
```

Figure 3. One CSV line from Question-answering/Basketball.csv file.

We followed the traditional question answering dataset structure proposed by [Rajpurkar et al. 2016]. This representation allows for efficient storage, retrieval, and analysis of the question record data. The context-question-answer triples provide a comprehensive record of the information in the dataset, enabling further exploration and utilization in various applications, such as training machine learning models and conducting data-driven research in natural language processing and information retrieval.

4. Analysis

In this section, we investigate the result of our work considering two different perspectives. In Section 4.1, we detail the characteristics of the wiki pages used to extract data from sports. In Section 4.2, we analyze the questions in QASports considering different types of "wh-words" queries.

4.1. Wiki Pages

The objective of analyzing the wiki pages used to extract data from sports is to gather information about the categories stored in each page. We obtained the following findings.

- For soccer, out of the total 22k wiki pages: (i) 5,987 (26%) are about players; (ii) 2,353 (10%) are about clubs; and (iii) 757 (3%) are about stadiums.
- For American football, out of the total 26k wikis: (i) 548 (2%) are about players; (ii) 921 (3%) about are about teams, with 792 being college teams; and (iii) 429 (2%) are about stadiums. Furthermore, most pages detail information and performance about team seasons.
- For basketball, out of the total 5k wikis: (i) 1,055 (20%) are about players; (ii) 164 (3%) are about arenas; and (iii) 1,220 (23%) are about teams. Among these teams, 50 are NBA teams, 71 are women teams, either from WNBA (Women's National Basketball Association) or college basketball.

Despite the vast number of wiki pages used in our work, we noted that there is a significant lack of information regarding women. Considering soccer, there are only 170 and 44 pages for female players and clubs, respectively. Furthermore, there is no information about women football. Fortunately, we also concluded that basketball wiki pages contain more information about women teams than the NBA.

Another useful analysis refers to the top categories from the extracted sports wiki pages. In descending order, we can cite players, teams, and stadiums, with the Football wiki being an exception, with the majority of pages being records of an individual season of a team, detailing matches and giving overall stats.

4.2. "Wh-words" Queries.

The objective of employing the traditional analysis of question answer datasets using "wh-words" [Nguyen et al. 2016] is to identify the information required to answer a question using QASports, as follows. A "Who" question is answered by the name of a person. A "Where" question determines a place. Furthermore, "What", "How", and "Why" questions have broader answers, usually requiring multiple sentences of reasoning.

Table 3 shows that majority of the questions are "What" and "How" types. Therefore, QASports requires complex reasoning from the QA model being trained and tested using these questions. Another aspect that makes a question harder for the QA algorithm is to have the words in the question different from the ones used to display the information in the context. For example, when the question uses the word "height" and the context says he "measured 6 feet 9.25 inches", and later also has the length of his wingspan, which is the same type of measurement for height, so it was not as simple as matching the expected type of information, in this case a measurement.

Table 3. Types of questions in QASports

What	How	Who	When	Where	Which	Why	*Total*
43%	23%	17%	8%	4%	3%	2%	100%

5. Scenarios for Using QASports

The QASports dataset can be used in several analytical applications to train and evaluate QA Systems. In this section, we describe scenarios for motivating potential dataset users to find creative use cases and solve issues commonly encountered in everyday life.

Sports news and analysis: A QA system trained on QASports can provide quick answers to sports news and analysis questions. For instance, a user can submit the query "What team won the NBA in 2022?".

Sports betting: A QA system using QASports can provide information about competing teams and betting values over time. For instance, a user can submit the queries "What were the results of the last five games between teams A and B?" and "What were the betting values for soccer and basketball last year?"

Player investments: A QA system can help evaluate a player by answering questions against QASports related to their past performances. For instance, a user can ask "What is the player's average score for the last two seasons?" and "How many baskets did the player hit last year?"

Tactical decision-making: Using QASports, the system can provide information related to the opposition team's playing style, weaknesses, and strengths to coaches and players. For instance, a coach can ask the following questions: "What is the preferred formation of the opposing team?" and "What are the weaknesses in defense of the opposing team?"

Sports trivia games: QASports can support the creation of trivia games and quizzes for sports fans. For instance, a user can ask about identifying the player who holds the record in the latest seasons.

Overall, a QA system using the QASports dataset can assist different users, including investors with player investments and online betting to decision-making coaches. Using a QA system, coaches can get accurate and timely information to make decisions during game matches. Furthermore, investors can gain a competitive advantage and quickly improve their analysis by gathering information.

6. Conclusion

In this paper, we introduced QASports, the *first* large sports question answering dataset for extractive answer questions. The QASports data is composed of three pieces. The first data consists of 54k pages from popular sports Fandom wikis, specifically soccer, American football, and basketball. The pages are stored in individual JSON files and collectively take up 11 GB. The next step was to extract sentences, called contexts, and store them together in a CSV file for each sport, resulting in 284k lines in 270 MB. In last, the data used for training and testing QA algorithms is 1.5 million sets of context, question, and answer. These are divided into three large CSV files, one for each sport, totaling 1.97 GB. Moreover, the wiki pages took about 40 hours to process and clean. Generating questions and answers took approximately 36 days of processing in an environment with 16 GB RAM, 500 GB HD, and an Intel i5 11th Gen processor with 6 cores.

We also investigated the wiki pages we used to extract sports data. We analyzed the questions into "wh-questions" categories to better understand the type of data needed to answer them. Moreover, we highlighted different ways to use QASports for knowledge extraction. We can use the information about players and teams to train and evaluate QA systems that help with decision-making and analysis. Even outside the scope of sports, this dataset can be used to train and evaluate both the Document Reader and the Document Retriever of general purpose QA systems. The QASports dataset[8] and the scripts[9] are publicly available for download and open for contributions and re-usability. To achieve state-of-the-art performance, we can train and fine-tune models with our dataset as our future work. We could also increase the range of question categories. In addition, we could expand our dataset by gathering data from many other sports.

Acknowledgements

We thank Amaris Consulting, São Paulo Research Foundation (FAPESP), Brazilian Federal Research Agency CNPq, and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil (CAPES) [Finance Code 001] for supporting this work. C. D. Aguiar has been supported by the grant #2018/22277-8, FAPESP. P. C. Jardim has been supported by the grant #2023/08293-9, FAPESP.

[8] QASports dataset - osf.io/n7r23, or huggingface.co/datasets/QASports

[9] Dataset scripts - github.com/leomaurodesenv/qasports-dataset-scripts

References

Alvares, J. C. M. and Ribeiro, M. R. (2019). Soccernews2018: a dataset of statistics and news of the 2018 brazilian soccer championship. In *XXXIV Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2019*, pages 440–446, Fortaleza, CE, Brazil. SBC.

Athira, P., Sreeja, M., and Reghuraj, P. (2013). Architecture of an ontology-based domain-specific natural language question answering system. *International Journal of Web & Semantic Technology*, 4(4): article number 31.

Bartolo, M., Roberts, A., Welbl, J., Riedel, S., and Stenetorp, P. (2020). Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Beal, R., Norman, T. J., and Ramchurn, S. D. (2019). Artificial intelligence for team sports: a survey. *The Knowledge Engineering Review*, 34:e28.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Liu, Q., Jiang, S., Wang, Y., and Li, S. (2020). LiveQA: A question answering dataset over sports live. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 1057–1067, Haikou, China. Chinese Information Processing Society of China.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mishra, A. and Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, 28(3):345–361.

Mittell, J. (2009). Sites of participation: Wiki fandom and the case of lostpedia. *Transformative Works and Cultures*, 3(3):1–10.

Moraes, L. M. P., Jardim, P., and Aguiar, C. D. (2023). Design principles and a software reference architecture for big data question answering systems. In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS)*, pages 57–67. INSTICC, SciTePress.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Pan, A., Chan, J. S., Zou, A., Li, N., Basart, S., Woodside, T., Ng, J., Zhang, H., Emmons, S., and Hendrycks, D. (2023). Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Ribeiro, M. R., Barioni, M. C. N., de Amo, S., Roncancio, C., and Labbé, C. (2017). Soccer2014ds: a dataset containing player events from the 2014 world cup. In *XXXII Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD 2017*, pages 278–285, Uberlândia, MG, Brazil. SBC.

Richardson, M., Burges, C. J., and Renshaw, E. (2013). MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.