

Resolving duplicates in Large Multiple-Choice Questions Repositories

Valentina Albano¹, Donatella Firmani², Luigi Laura³, Jerin George Mathew², Anna Lucia Paoletti¹ and Irene Torrente⁴

¹Dip. Funzione Pubblica, Corso Vittorio Emanuele II, 116, 00186 Rome, Italy.

²Sapienza University, Piazzale Aldo Moro, 5, 00185 Rome, Italy

³Uninettuno University, Corso Vittorio Emanuele II, 39, 00186 Rome, Italy

⁴Formez, Viale Marx, 15, 00137 Rome, Italy

Abstract

Multiple-choice questions (MCQs) are commonly used in educational assessments and professional certification examinations. However, managing vast collections of MCQs presents numerous challenges, including maintaining their quality and relevance. A notable issue in such repositories is the occurrence of conceptually identical questions presented in varied forms. These duplicates, while different in wording, fail to enhance the value of the repository. In this extended abstract, we present our approach for identifying and handling potential duplicate questions in large MCQ databases. Our proposed method involves three primary stages: initial pre-processing of MCQs, calculation of similarity based on Natural Language Processing (NLP) techniques, and a graph-based method for exploring these similarities.

Keywords

multiple-choice questions, entity resolution, record linkage, graph communities

1. Introduction

Multiple-Choice Questions (MCQs) are widely utilized for knowledge assessment across various domains, from university admissions and job evaluations to self-assessment and entertainment, including popular game shows and mobile gaming apps. Large-scale standardized tests typically feature MCQs with four response options: one correct answer and three distractors.

Academic research primarily focuses on the effectiveness of MCQs as evaluation tools. Azevedo, Oliveria, and Damas Beites' study [1] is exemplary in exploring methods for fair student assessments through MCQ analysis. Learning Analytics, as defined in [2], involves the comprehensive measurement and analysis of learner data to optimize educational environments. Furthermore, the laborious nature of manual MCQ creation has led researchers to the development of automatic generation techniques. These range from using resources like WordNet and shallow parsing to advanced methods involving ontologies and deep neural networks [3].

IRCDL 2024 20th conference on Information and Research science Connecting to Digital and Library science formerly the Italian Research Conference on Digital Libraries Bressanone, Brixen, Italy - 22-23 February 2024


✉ V.Albano@governo.it (V. Albano); donatella.firmani@uniroma1.it (D. Firmani);

luigi.laura@uninettunouniversity.net (L. Laura); mathew@diag.uniroma1.it (J. G. Mathew);

A.Paoletti@funzionepubblica.it (A. L. Paoletti); itorrente@formez.it (I. Torrente)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Unlike previous research, our approach focuses on the maintenance of large Multiple-Choice Questions (MCQs) repositories with a data quality perspective. Specifically, we focus on the identification of conceptually similar questions within these repositories. Our method can be used by MCQs database administrators to identify overly similar questions and verifying question coherence with syllabus areas and coverage. Our intuition is rooted in the belief that redundant or less coherent questions can be accumulated in MCQs repositories over time – e.g., upon addition and merging operation – which can dilute their quality and effectiveness.

The primary feature of our method is the utilization of recent Natural Language Processing (NLP) techniques in an entirely unsupervised environment. These advanced NLP methods, based on the popular Transformer architecture, are capable of identifying a substantial number of relevant matches. However, they also carry the risk of generating false positives, where questions might be inaccurately classified as similar. To mitigate this issue of false positives, our approach incorporates an innovative graph exploration technique. This technique focuses on identifying candidate matches from questions that are part of densely connected graph communities. By doing so, our method not only minimizes the likelihood of false positives but can also uncover the underlying structure of the MCQs repository.

Outline. Section 2 describes the main components of our approach and a case study from a large-scale training project by the Italian Government. Section 3 summarizes our experimental results in the main application scenarios. Section 4 describes related works and Section 5 presents conclusive remarks. For more detailed insights, please refer to the recent journal paper [4] and to the conference paper in [5] presenting an earlier version of our approach.

2. Overview

In this section, we outline our proposed workflow for managing large Multiple Choice Questions (MCQs) repositories, which includes the following four key steps:

1. **Similarity Computation:** Compute similarity scores between all pairs of questions in an unsupervised manner, without the need for a labeled dataset.
2. **Threshold Definition:** Establish a similarity threshold, σ , to preliminarily distinguish between similar and dissimilar questions. This threshold is adjustable, allowing interactive user selection.
3. **Graph Construction:** Construct a graph, G_σ , where nodes represent questions, and edges connect questions with similarity scores above the threshold, σ .
4. **Graph Exploration:** Employ graph visual analysis, focusing on graph communities to identify clusters of potentially similar questions.

Case study. We demonstrate our approach using the MCQs database from the *Competenze Digitali* program.¹ The program aims to provide non-IT public employees with personalized e-learning training in basic digital skills. Key elements include (i) a Syllabus outlining minimum digital skills required for public employees, (ii) an online platform for skill gap analysis, training course definition, and delivery, and (iii) a catalog of quality training developed in collaboration

¹Database access is restricted, but the new program is publicly available at <https://www.syllabus.gov.it>.

Question: What is the definition of “cookies”?
A. Small files that store information about online browsing on the user’s computer or device.
B. Small files that track and collect user’s online behaviors for targeted advertising purposes.
C. Tokens generated by web applications to authenticate user sessions and enable personalized features.
D. Privacy settings that allow users to control the information shared with websites they visit.

Figure 1: Sample question and answers inspired by a similar question in the MCQ Dataset.

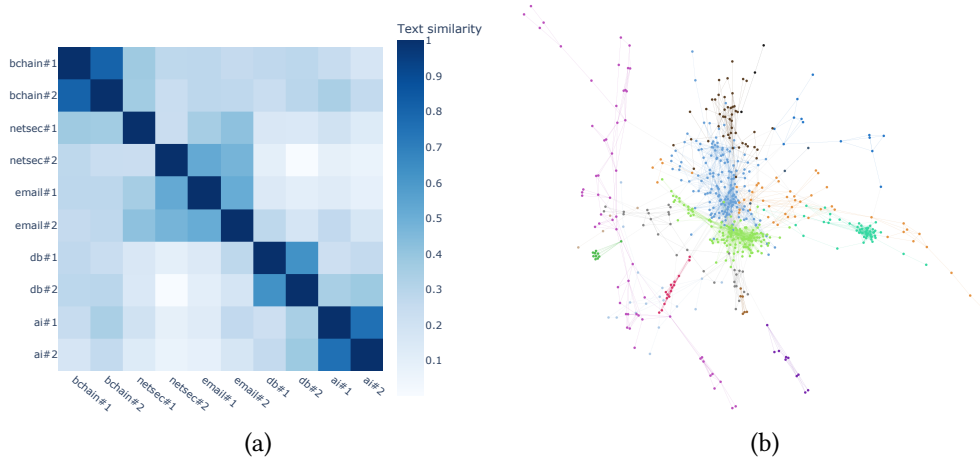


Figure 2: (Left) Similarity matrix produced by ST-MPNet. We indicate for each question its corresponding topic, for instance ai#1 and ai#2 are two sentences related to AI. **(Right)** The largest CC in $G_{0.7}$ (ST-MPNet) with node colors highlighting the community structure.

with major public and private entities. The MCQ dataset of the *Competenze Digitali* program comprises 798 Italian language questions, each presenting four candidate answers, of which only one is correct. Every question corresponds to a particular syllabus, which groups together questions related to the same topic (e.g. computer networks). In total, there are 11 distinct syllabi in the dataset. We provide a sample question in Figure 1.

3. Applications

3.1. Identifying redundant questions

We describe a method for identifying overly similar questions in the *Competenze Digitali* MCQs database. In our experiment, we tested four multilingual models from the Sentence Transformers library [6] – specifically fine-tuned with Italian data – that are ST-Roberta, ST-DistilUSE, ST-MiniLM, and ST-MPNet. We assessed their performance on a subset of our MCQ dataset, which included manually selected question pairs related to the same syllabus topic, and identified ST-MPNet as the best-performing model. The resulting similarity matrix from ST-MPNet, illustrated in Figure 2a, displays a pattern of darker 2×2 squares along the diagonal, indicating accurate semantic matches of pairs of most similar questions.

Question	Syllabus area	Similarity
Q-1.1.1.1-1	S1.1	0.415891
Q-1.1.1.1-1	S1.2	0.147890
Q-1.1.1.1-1	S1.3	0.254660
Q-1.1.1.1-1	S2.1	0.158570
Q-1.1.1.1-1	S2.2	0.202694
...
Q-5.2.3.5-8	S3.2	0.041329
Q-5.2.3.5-8	S4.1	0.096101
Q-5.2.3.5-8	S4.2	0.073620
Q-5.2.3.5-8	S5.1	0.106045
Q-5.2.3.5-8	S5.2	0.175479

Figure 3: The similarity values between the questions and the areas of the Syllabus.

A critical consideration in our approach is whether to compute similarity using only the question text, the question with the correct answer, or the question with all answer options. Our analysis suggests that including all answer options provides the most accurate results, particularly for questions with identical wording, such as those in our repository that simply ask, “Which of the following statements is false?”.

When analyzing the similarity values, we found a spectrum ranging from completely unrelated questions (similarity 0) to those that were suspiciously similar (similarity values between 0.98 and 0.99). To address the intermediate range of similarities, we implemented a graph-based method by setting a threshold $\sigma = 0.7$. We examined the graph $G_{0.7}$, consisting of 718 nodes and 3,836 edges, and applied the Clauset-Newman-Moore algorithm [7] for community detection.² The communities identified, as depicted in Figure 2b, are often granular enough to enable manual inspection and identification of similar question groups. For cases requiring additional analysis, node-centrality methods such as [8, 9] can be employed to enhance visual exploration.

3.2. Verifying syllabus coherence

Using the previously described similarity techniques and a structured syllabus, we can efficiently assess the alignment of questions with specific syllabus areas. For instance, in the Competenze Digitali program, the syllabus comprises five main areas and 11 sub-areas. Each question in our database is uniquely identified by the area and sub-area numbers it corresponds to.

This structure allows us to analyze the similarity between each question and its designated sub-area within the syllabus. In our dataset we have 798 questions and 11 sub-areas, so we computed $798 \cdot 11 = 8778$ values of similarity. In Table 3 we report the head and the tail of the values; the first items are the similarity score of the first question in the dataset against the first sub-areas of the Syllabus. It is easy to see that the largest score, among the ones shown in the table, is exactly for the sub-area of the Syllabus it belongs, i.e., Q-1.1.1.1-1 belongs to S1.1. The same happens for the last question, i.e., Q-5.2.3.5-8 belongs to S5.2. About half of the questions in our dataset showed the highest similarity with their respective syllabus sub-areas.

²The results of a baseline approach based on connected components instead of communities are available in [5].

This outcome highlights the effectiveness of our approach in ensuring that questions are relevant and aligned with specific curriculum areas. By examining these similarity scores, educators and administrators of the MCQs repository can identify syllabus areas needing more focus or refinement and detect potential redundancies in their question sets.

4. Related Works

The exploration of duplicate question detection has been investigated in the Q&A domain, aiming to efficiently answer queries by linking to similar, previously answered questions in Q&A forums. Notable recent works include those of Li et al. [10], who focused on medical Q&A platforms, and Kamienski et al. [11], who concentrated on game development forums. Both studies developed deep-learning systems for recognizing similar questions. Li et al. trained a Long-Short Term Memory (LSTM) neural network on pairs of questions, aligning semantically similar queries within the LSTM's vector space. Kamienski et al. combined large pre-trained deep learning models with supervised techniques, using features from models like MPNet to train a supervised model that predicts similarity scores between sentence pairs.

In the field of learning analytics, there has been limited investigation into identifying duplicate questions. A notable study is presented in [12], which developed a machine learning-based system for managing large question paper databases. It trains an XGBoost model [13] on manually-selected features like structural attributes and word embeddings, using labeled duplicate question pairs from Quora to identify semantically similar English sentence pairs.

Unlike these methods, our study leverages pre-trained large language models in an unsupervised manner, without requiring ground-truth labels. Additionally, our approach incorporates a graph construction phase to facilitate the identification of duplicates.

Other works. The general problem of identifying duplicate records in databases is known in literature as Entity Resolution. Data management applications typically use supervised methods [14, 15] and external knowledge bases [16] to learn vector representations of records and attributes. In this paper, we focus instead on unsupervised similarity computation and leave the final decision to a human expert, in the spirit of oracle-based approaches such as [17, 18].

5. Conclusive remarks

We describe our approach for the maintenance of large Multiple-Choice Questions (MCQs) repositories, based on our experience with a large-scale training project by the Italian Government. Our method employs Natural Language Processing (NLP) techniques, particularly *Transformer* architectures, to semantically detect similar pairs of questions in MCQs repositories, aiming to go beyond traditional word co-occurrence analysis while acknowledging the potential for false positives. To address these false positives, we incorporate a graph exploration strategy that focuses on community structures within the *similarity graph*, thus enhancing the precision of similarity assessments by analyzing relationships within the same community.

Future works include exploring algorithms for efficiently merge multiple MCQ repositories and developing automated tools for aligning questions upon syllabus updates, thereby reducing the manual effort required for repository maintenance.

Acknowledgements

The authors have been partially supported by SEED PNR Project “FLOWER” “Frontiers in Linking records: knOWledge graphs, Explainability and tempoRal data” and Sapienza Research Project B83C22007180001 “Trustworthy Technologies for Augmenting Knowledge Graphs”.

References

- [1] J. M. Azevedo, E. P. Oliveira, P. Damas Beites, Using learning analytics to evaluate the quality of multiple-choice questions: A perspective with classical test theory and item response theory, *The International Journal of Information and Learning Technology* 36 (2019) 322–341. doi:10.1108/IJILT-02-2019-0023.
- [2] P. Long, G. Siemens, G. Conole, D. Gasevic (Eds.), *Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK 2011, Banff, AB, Canada, February 27 - March 01, 2011*, ACM, 2011.
- [3] R. Mitkov, H. Le An, N. Karamanis, A computer-aided environment for generating multiple-choice test items, *Natural language engineering* 12 (2006) 177–194. doi:10.1017/S1351324906004177.
- [4] V. Albano, D. Firmani, L. Laura, J. G. Mathew, A. L. Paoletti, I. Torrente, Nlp-based management of large multiple-choice test item repositories, *Journal of Learning Analytics* (2023) 1–16.
- [5] V. Albano, D. Firmani, L. Laura, A. L. Paoletti, I. Torrente, Managing large multiple-choice test item repositories, in: *Proceedings of the 26 International Conference Information Visualisation (IV22)*, 2022. doi:10.1109/IV56949.2022.00054.
- [6] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2019. URL: <http://arxiv.org/abs/1908.10084>. doi:10.48550/arXiv.1908.10084.
- [7] A. Clauset, M. E. J. Newman, C. Moore, Finding community structure in very large networks, *Physical Review E* 70 (2004). doi:10.1103/physreve.70.066111.
- [8] G. Ausiello, D. Firmani, L. Laura, The (betweenness) centrality of critical nodes and network cores, in: *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)*, IEEE, 2013, pp. 90–95. doi:10.1109/IWCMC.2013.6583540.
- [9] G. Ausiello, D. Firmani, L. Laura, Real-time monitoring of undirected networks: Articulation points, bridges, and connected and biconnected components, *Networks* 59 (2012) 275–288. doi:10.1002/net.21450.
- [10] Y. Li, L. Yao, N. Du, J. Gao, Q. Li, C. Meng, C. Zhang, W. Fan, Finding similar medical questions from question answering websites, 2018. doi:10.48550/arXiv.1810.05983. arXiv:1810.05983.
- [11] A. Kamienski, A. Hindle, C.-P. Bezemer, Analyzing techniques for duplicate question detection on q&a websites for game developers, *Empirical Software Engineering* 28 (2023) 17. doi:10.1007/s10664-022-10256-w.
- [12] S. Mukherjee, N. S. Kumar, Duplicate question management and answer verification

- system, in: 2019 IEEE Tenth International Conference on Technology for Education (T4E), 2019, pp. 266–267. doi:10.1109/T4E.2019.00067.
- [13] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.
- [14] U. Brunner, K. Stockinger, Entity matching with transformer architectures-a step forward in data integration, in: EDBT 2020, 2020. doi:10.5441/002/edbt.2020.58.
- [15] M. Ebraheem, S. Thirumuruganathan, S. R. Joty, M. Ouzzani, N. Tang, Distributed representations of tuples for entity resolution, Proc. VLDB Endow. 11 (2018) 1454–1467. URL: <http://www.vldb.org/pvldb/vol11/p1454-ebraheem.pdf>. doi:10.14778/3236187.3236198.
- [16] A. S. Andreou, D. Firmani, J. G. Mathew, M. Mecella, M. Pingos, Using knowledge graphs for record linkage: Challenges and opportunities, in: M. Ruiz, P. Soffer (Eds.), Advanced Information Systems Engineering Workshops - CAiSE 2023 International Workshops, Zaragoza, Spain, June 12-16, 2023, Proceedings, volume 482 of *Lecture Notes in Business Information Processing*, Springer, 2023, pp. 145–151. URL: https://doi.org/10.1007/978-3-031-34985-0_15. doi:10.1007/978-3-031-34985-0_15.
- [17] D. Firmani, S. Galhotra, B. Saha, D. Srivastava, Robust entity resolution using a crowdoracle, IEEE Data Eng. Bull. 41 (2018) 91–103. URL: <http://sites.computer.org/debull/A18june/p91.pdf>.
- [18] S. Galhotra, D. Firmani, B. Saha, D. Srivastava, Efficient and effective er with progressive blocking, The VLDB Journal 30 (2021) 537–557. doi:10.1007/s00778-021-00656-7.