Near-duplicate Question Detection

Preetam Prabhu Srikar Dammu*
University of Washington
preetams@uw.edu

Omar Alonso
Amazon
omralon@amazon.com

ABSTRACT

Suggesting relevant questions to users is an important task in various applications, such as community Q&A or e-commerce websites. To ensure that there is no redundancy in the selected set of candidate questions, it is essential to filter out any near-duplicate questions. Identifying near-duplicate questions has another use case in light of the adoption of Large Language Models (LLMs) – fetching pre-computed answers for similar questions. However, identifying the similarity of questions is a bit more complex in comparison to generic text, as questions entail open-ended information that is not explicitly contained within the wording of the question itself. We introduce a taxonomy that accounts for the subtle intricacies characteristic of near-duplicate questions and propose a method for detecting them utilizing the capabilities of LLMs.

ACM Reference Format:
Preetam Prabhu Srikar Dammu and Omar Alonso. 2024. Near-duplicate Question Detection. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Users often pose the same question using different words or phrases, and in applications with a multitude of users, this could lead to significant redundancy due to duplication. Identifying near-duplicate questions is an effective strategy for organizing content and enables efficient processing for downstream tasks.

Identifying near-duplicate questions is more challenging than exact duplicates due to subjectivity. For example, `Q1: Which laptop is the best under $3000?` and `Q2: Which is the best laptop under $3000?` are an obvious duplicate pair. On the other hand, `Q1: Which laptop is the best and costs less than $3000?` and `Q2: What is the best laptop under 3k for heavy users?` can be considered as near-duplicates since subtle differences exist, yet both questions are essentially seeking similar information.

Various methods, ranging from lexical matching to semantic similarity on text embeddings, effectively quantify text similarity in web pages, documents, and statements. However, we show that they have limited applicability when it comes to questions.

Removing stop-words is a standard pre-processing step, but for questions, *wh question words*, typically seen as stop-words, are vital for conveying intent. On the other hand, sentence embedding techniques such as SBERT [9] aim to capture the meaning of a given sentence, yet they are highly influenced by keywords.

Another challenge with questions is that they are typically short. This hindrance is also observed in the identification of near-duplicate tweets, where the authors of [11] relied on supplementary features like user similarity and information extracted from hyperlinks. In a similar fashion, [7] attempt to use answers in their near-duplicate question detection method. However, this may not be practical in use cases where answers to the questions are not readily available.

As the notion of similarity between questions is associated with the kind of responses they invoke, it is insufficient to rely on simple semantic or lexical matching of the question texts which do not completely account for such information. For example, `Q1: Why should I upgrade to iPhone 14?` and `Q2: When should I upgrade to iPhone 14?` might generate a high similarity score as there are many overlapping words, but they will invoke different answers as they have different intents. Yet, traditional methods generate high similarity scores for these questions, as shown in Figure 1.

Utilizing answers to identify duplicate question pairs may seem logical but it is problematic, as different questions can have similar valid answers and vice versa. For instance, `Q1: Does this watch need batteries?` and `Q2: Does this bike need an external lock?` both could be correctly answered with "No, it does not.". This makes answers, especially generic ones, an ambiguous signal. It's better to have DQI methods not rely on answers, avoiding the computational burden of generating and processing answers, particularly in cost-intensive LLM-based applications.

In this paper, we present a taxonomy for near-duplicate questions and introduce Zero-shot near-Duplicate Questions Identifier (ZSDQI), a new method capable of detecting a wider range of near-duplicate questions using zero-shot learning.

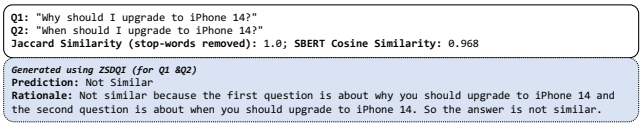

Figure 1: Intent detection & reasoning capabilities of ZSDQI.

2 RELATED WORK

Duplicate questions identification (DQI) is a well-studied problem with various methods proposed in the literature [7, 13, 14]. These can be widely categorized into two common approaches, which are *sentence encoding* and *sentence interaction-aggregation* [5]. Sentence encoding methods are broadly applicable, useful for tasks like clustering or document retrieval, while interaction-aggregation models are specific to sentence pair tasks like paraphrase identification, similarity detection, or natural language inference.

Embedding models are trained to project inputs to an embedding space where semantically similar instances are close to each other.

*Work done during internship at Amazon.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2024 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Category	Suitable Method(s)	Question 2	JS	ES	ZSDQI
Lexical	Lexical; Semantic	Battery on the iPhone 14 is decent?	1.0	0.96	Prediction: Similar
Paraphrased	Lexical; Semantic	How acceptable is the battery on the iPhone 14?	0.6	0.97	Prediction: Similar
Informational	LLMs	Does the iPhone 14 last long enough on a single charge?	0.25	0.74	Prediction: Similar
Evidence-based	LLMs + Knowledge Aug.	How good is the battery on the best-selling smartphone?	0.12	0.80	Prediction: Not similar

Table 1: Near-duplicate question categorization example: `Is the battery on the iPhone decent?` JS (Jaccard with stopwords removed), ES (Embedding similarity with cosine similarity between SBERT embedding pairs), ZSDQI.

This approach performs well in identifying dissimilar questions when they are on different topics, however, falls short when they are on the same topic. Such behaviour is observed as encoding models do not adequately capture the asking emphasis of questions [14]. In [5], the authors find that accounting for inter-sentence interactions leads to an improved performance. However, in cases where reasoning is required for judging if two questions are similar, LLMs are a more suitable option as they have been shown to possess such capabilities [4, 12] while the prior works have not.

3 CATEGORIZATION

We define question similarity as follows: *"If two questions with similar intent seek similar information about the same entity, they can be considered as near-duplicates."* This interpretation of similarity could be more suitable for identifying near-duplicate question pairs, as it examines the similarities of questions by anticipating the similarities of their potential responses. It eliminates the dependence on the actual answer text when evaluating the similarities between the questions. However, it is not straightforward to operationalize this definition, as it entails multiple complex sub-tasks, such as text comprehension, intent detection, entity recognition, and reasoning. Our experiments demonstrate that LLMs can perform all these sub-tasks, a capability not observed in previous DQI methods.

Detecting question intent is important as it determines the answer properties. Recognizing entities is an essential step for subsequent reasoning, and we show that LLMs identify and use internal knowledge associated with detected entities. Text comprehension and reasoning are crucial for evaluating if two questions are seeking similar information, even when worded differently.

We propose the following categorization of near-duplicate questions based on their characteristics and identification complexity. We arrived at four categories after considering the identification method's technical properties required for reliably detecting near-duplicates. While semantic capabilities may be enough for lexical and paraphrased duplicates, advanced capabilities such as entity recognition and reasoning are required for methods suitable for handling informational and evidence-based duplicates.

Lexical. These instances have a significant number of overlapping words. Methods that rely on word counts may deliver acceptable performance. Jaccard similarity expressed as $J(Q_1, Q_2) = (|Q_1 \cap Q_2|)/(|Q_1 \cup Q_2|)$ is used, where Q_1 is the set of words in the first question and Q_2 is the set of words in the second question.

Paraphrased. These are instances that are paraphrased but rather easy to spot without requiring much reasoning. Semantic similarity methods could be sufficient to identify such duplicates. Typically, the similarity score is calculated as $d(e_{Q1}, e_{Q2})$, where d is any distance measure, and e_{Q1}, e_{Q2} are embeddings of Q_1 and Q_2 extracted using any text embedding methods [8–10].

Informational. Questions that seek similar information about the same topic or entity are considered as informational duplicates. Instances that may not be direct paraphrases of each other, yet invoke similar responses, fall in this category. For example, `Q2: Does the iPhone 14 last long enough on a single charge?` does not consist important keywords such as "battery" or synonyms of "decent" which are present in `Q1: Is the battery on the iPhone 14 decent?`, yet these questions are essentially asking for the same information.

Evidence-based. These instances require evidence in order to be accurately judged as duplicates or otherwise. They could be questions whose answers might change over time (like the best-selling smartphone model) or not common knowledge (like the storage capacity of a less-known smartphone model). For example, in Table 1, the question `Q2: How good is the battery on the best-selling smartphone?` is listed under this category as we would only be able to judge if Q1 and Q2 are duplicates if we had a dynamic and reliable knowledge base that could confirm if iPhone 14 is the best-selling smartphone. Although, for the given example, the proposed solution utilizes its internal knowledge base to arrive at a prediction which may or not be correct. In Section 7, we discuss future work that could potentially enable accurate processing of evidence-based duplicates.

Table 1 presents examples for each category, comparing Jaccard similarity, embedding similarity, and predictions from our method. Lexical methods like Jaccard struggle with complex categories by penalizing differently worded similar questions. Embedding similarity scores high for same-topic questions but fails in detecting subtle differences in dissimilar questions, as shown in Figure 1. Our method, however, provides predictions with intuitive justification.

4 PROPOSED METHOD

Our method uses LLMs to perform text comprehension, entity recognition, intent detection, and reasoning through zero-shot learning. This can be achieved by utilizing a suitable LLM model and querying it with zero-shot chain-of-thought prompting [4].

Previous work has shown that using intermediate reasoning steps while interacting with an LLM helps the model to perform better in tasks that require complex reasoning [12]. This is crucial for our task, as reasoning is an unavoidable sub-task for identifying informational duplicates as discussed in Section 3.

Pipeline Details. The algorithmic details of the proposed solution are presented in Algorithm 1. Certain components of the method remain static during the inference phase, such as the LLM model and the query templates. In our experiments, we use the FLAN T5-XXL [1] model, but any LLM of choice can be used. The prompts issued

Algorithm 1 ZSDQI

Input: Q_1: question 1; Q_2: question 2;
Static: *model*: LLM; $template_R$: reasoning template;
$template_{LE}$: label extraction template;
Operations: $\textbf{prompt}_{\textbf{Gen}}()$: populates template with inputs;
Output: *prediction*, *rationale*

1: **procedure** ZSDQI(Q_1, Q_2)
2: $reasoning\_prompt \leftarrow prompt_{Gen}(template_R, Q1, Q2)$
3: $rationale \leftarrow model(reasoning\_prompt)$
4: $extraction\_prompt \leftarrow prompt_{Gen}(template_{LE}, rationale)$
5: $label \leftarrow model(extraction\_prompt)$
6: **return** $(label, rationale)$
7: **end procedure**

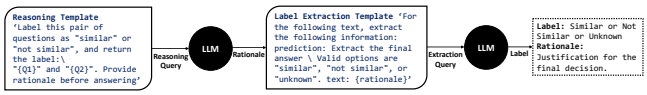

Figure 2: Pipeline of ZSDQI. Two sequential LLM invocations are made to arrive at a prediction supported by a justification.

to the LLM model are generated on runtime by the $prompt_{Gen}()$ function, which populates the template with the inputs. The algorithm takes the question pair as input and return the label and its justifying rationale. As shown in Figure 2, we issue two sequential queries to the LLM, the first one to generate the reasoning along with the answer and the second to extract the label. The prompt templates for both queries are shown in Figure 2.

Reasoning Template. In the first query to the model, we ask the LLM to generate rationale as to why a pair of questions would be considered as duplicates before arriving at the final decision. This invokes the LLM's reasoning capabilities [4, 12]. We instruct the model to either label the pair as similar or not similar.

Label Extraction Template. The second query to the LLM is based on the response from the first query, and is used for extracting the final label. As LLMs tend to generate answers in natural language with varying sentence structures, a second query is required for extracting only the final prediction.

5 EXPERIMENTS AND RESULTS

We use the Quora Questions Pairs (QQP) dataset [3] for evaluation. Previous works on this dataset have used different test splits, and in order to make fair comparisons, we use corresponding test splits for comparing performance with our method.

The first column of Table 2 shows a 13.42% performance increase over methods in [10], which are generic architectures with broad applications. In contrast, most of the methods listed in the second column are *sentence interaction-aggregation models* [5] which are mostly suitable for sentence pair modeling tasks. LLMs, as foundation models, have a wider range of applications and capabilities. A comprehensive list of previous DQI methods is available in [5].

While some methods in Table 2 report higher accuracy on QQP 10K test split, it is worth noting that they were also trained on QQP. Notably, ZSDQI has not been trained or finetuned for QQP, or any dataset for the given task. Hence, making the proposed solution *dataset agnostic*. This is essential for deployment, allowing plug-and-play use across various domains without fine-tuning. Moreover, significant mislabeling in QQP may affect reported scores [5, 6].

QQP 6K		QQP 10K	
Model	**Accuracy**	**Model**	**Accuracy**
Jcrd	69.53	ABCNN	63.59
SVM-bas	64.93	PWIM	83.40
SVM-adv	68.56	pt-DecAtt$_{char}$	87.54
CNN	59.90	pt-DecAtt$_{word}$	88.40
DNN	69.53	MFAE (ELMo)	89.61
DCNN	71.48	**MFAE (BERT Ensemble)**	**90.54**
ZSDQI	**84.90**	ZSDQI	84.35

Table 2: Performance comparison on QQP 6K test split [10] and QQP 10K test split [14]. None of the previous methods listed in this table are *dataset agnostic*, except for ZSDQI.

Accuracy is one among many considerations when it comes to deployment, as there are other desirable characteristics which improve the utility of a DQI method – such as generalizability and robustness to distribution shifts. Extensive finetuning to achieve high scores on test sets may lead to models that do not generalize well on other distributions. Distribution shifts are a cause for concern associated with ML models finetuned for specific datasets while dataset agnostic methods are more resistant towards them.

As most applications concerned with the near-duplicate question detection task are not limited to a particular domain, it is paramount to have a reliable solution that can perform equally well on a wide range of data distributions. For instance, the increasing popularity of personal chat agents would result in users issuing a wide range of questions to the same application, ranging from common sense questions to ones that require domain knowledge. In order to reliably identify duplicate questions, it is essential that the DQI method is dataset agnostic. ZSDQI benefits from the generalization capabilities of LLMs as they are trained on web-scale datasets.

The high out-of-the-box performance of the LLMs could be attributed to their inherent capabilities of text comprehension, entity recognition, intent detection, and reasoning. While most deep learning methods are good at text comprehension, they fall short in other areas. MFAE [14] is the only existing DQI method that considers asking emphasis or question intent. On the other hand, LLMs demonstrate all of these capabilities, and we show this for examples from the Quora dataset in Table 3.

Interestingly, LLMs are able to inherently identify entities and perform reasoning to arrive at decisions. This property is demonstrated clearly in the third row in Table 3, as the rationale correctly points out that the questions are about different subjects (trading and physics) by processing terms and entities such as Angel Broking, Einstein, and gravity. Reasoning capabilities are also demonstrated in row 4, as the rationale quotes the difference in price range would render the questions as different. Lexical or semantic methods may not be suitable for such cases, as the only difference in those questions is the number. In row 5, even when multiple names are present in the question text, the LLM successfully identified the person of focus and made the prediction accordingly. In row 6, the LLM accurately picks up the difference in location and presents it as the basis for predicting the question pair is dissimilar. And in row 2, we notice that the LLM is able to detect that the two questions are seeking for the same information, even they are asked differently. In Figure 1, the LLM also demonstrated capabilities of distinguishing between question intents.

	Question 1	Question 2	True	Pred	Generated Rationale
1	How did Hitler come to power?	Who followed Hitler to power?	NS	NS	Not similar because the first question is about how Hitler came to power and the second question is about who followed him to power.
2	Can we donate blood after getting a tattoo?	Can I donate blood if I have a tattoo?	S	S	Similar questions are asking if you can donate blood after getting a tattoo. The answer is similar.
3	Is it better to do day trading with Fyers or Angel Broking?	Why was Einstein so interested in Gravity right after Special Relativity?	NS	NS	Not similar because the questions are about different subjects. The first is about trading and the second is about physics.
4	Which is the best phone to buy under 2000?	Which is best phone to buy under 15,000?	NS	NS	Not similar because the price range is different. The first question is about a phone under 2000 and the second is about a phone under 15,000. The answer is not similar.
5	Why Cyrus Mistry has been removed from Tata Group?	Why was Cyrus Mistry sacked by Ratan Tata from Tata Sons?	S	S	Similar questions are about the same person. The answer is similar.
6	What is a good solar panel installation provider in Mar Vista , California CA?	Which is a good solar panel installation provider in Los Altos , California CA?	NS	NS	Not similar because the location is different. The first location is in Mar Vista and the second location is in Los Altos. The answer: not similar.

Table 3: Examples from QQP dataset. NS: Not Similar, S: Similar, True: Dataset Label, Pred: Predictions by ZDSQI, Generated Rationale: Justification generated by ZSDQI.

6 DISCUSSION AND FUTURE WORK

In addition to its performance, there are several motivating factors for using the proposed solution. One of them is employing ZSDQI as a teacher model in the absence of high-quality labeled data. Despite the presence of publicly available datasets for duplicate questions, they do not exhaustively represent all domains. Manually labeling data is often expensive and hard to obtain. In such cases, the proposed LLM solution would be suitable as it can be used without any training or finetuning.

Another advantage of ZSDQI is that it generates a justification for every prediction. As judging near-duplicates is a subjective task that may have debatable predictions, these explanations can be very informative for understanding model behavior and facilitating subsequent modifications. Changes to the model behavior could be achieved through simple modifications in the prompt, by specifying details that align with the desired outcomes. If further changes are required, advanced finetuning approaches like LoRA [2] could be used given that a small high-quality dataset is available.

Compared to existing techniques, ZSDQI is better equipped to handle informational duplicates. Since such near-duplicate pairs are not prevalent in the existing benchmark datasets, previous methods might report high scores on them. However, in real-world data, a substantial percentage of informational duplicates can be observed, particularly in applications with a large user base.

As ZSDQI is not tied to a specific LLM architecture, the LLM component can easily be swapped with a more suitable or powerful one whenever required. This flexibility is desirable, especially due to the fast-evolving nature of the field.

7 CONCLUSION

In this work, we propose a new method for identifying near-duplicate questions by operationalizing a comprehensive question similarity definition. To the best of our knowledge, our method is the first of its kind to leverage the emergent capabilities of LLMs to perform DQI. As such, this work opens a new path to the exploration of contemporary strategies for DQI. We also introduce a taxonomy for categorizing near-duplicate questions based on their characteristics, and discuss approaches suitable for detecting them. The proposed approach can serve as a base solution that could be enhanced in order to cater to broader requirements. Incorporating external knowledge would be a good strategy to support the identification of evidence-based duplicates. In order to minimize inference costs, LLM quantization could be a path worth exploring.

REFERENCES

[1] H. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, M. Wang, X.and Dehghani, S. Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

[2] E. Hu, Y. Shen, P. Wallis, Y. Allen-Zhu, Z.and Li, S. Wang, L. Wang, and W. Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[3] Shankar Iyer, Nikhil Dandekar, and Kornl Csernai. [n. d.]. First Quora Dataset Release: Question Pairs. https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

[4] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[5] W. Lan and W. Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. *arXiv preprint arXiv:1806.04330* (2018).

[6] H. T Le, D. Cao, T. Bui, L. Luong, and H. Nguyen. 2021. Improve Quora Question Pair Dataset for Question Similarity Task. In *Proc. of* RIVF. IEEE, 1–5.

[7] D. Liang, F. Zhang, W. Zhang, Q. Zhang, J. Fu, M. Peng, T. Gui, and X. Huang. 2019. Adaptive multi-attention network incorporating answer information for duplicate question detection. In *Proc. of* SIGIR. 95–104.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[9] N. Reimers and I. Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[10] J. Rodrigues, C. Saedi, V. Maraev, J. Silva, and A. Branco. 2017. Ways of asking and replying in duplicate question detection. In *Proc. of* SEM. 262–270.

[11] K. Tao, F. Abel, C. Hauff, G. Houben, and U. Gadiraju. 2013. Groundhog day: near-duplicate detection on twitter. In *Proc. of* WWW. 1273–1284.

[12] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, D. Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[13] W. Yin, H. Schütze, B. Xiang, and B. Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for computational linguistics* 4 (2016), 259–272.

[14] R. Zhang, Q. Zhou, B. Wu, W. Li, and T. Mo. 2020. What do questions exactly ask? mfae: Duplicate question identification with multi-fusion asking emphasis. In *Proc. of* SIAM. 226–234.