Evaluation Dimensions for Assessing Question Answer Systems for Lay Users: The Case of DiseaseGuru

Prakash Chandra Sukhwal, Atreyi Kankanhalli, Vaibhav Rajan

National University of Singapore
prakashs@nus.edu.sg, atreyi@comp.nus.edu.sg, vaibhav.rajan@nus.edu.sg

Abstract

Question answer (QA) systems can serve as vital tools to address lay users' information needs in healthcare. While QA systems have the potential to lessen information overload and provide quality answers to users, it is important to holistically evaluate their performance. Here we propose multiple dimensions for this purpose comprising lexical similarity, semantic similarity, absence of contradictions and readability of responses. We then use the dimensions to evaluate DiseaseGuru, a generative large language model-based chronic disease QA system we developed that integrates knowledge graph technology to provide quality responses to lay users. The results are presented comparing it with three benchmark algorithms across the different dimensions. We also propose metrics for lay users and medical professionals for a future field study to evaluate the system.

Introduction

In recent times, the Internet has become a popular source for people to seek information. Of these, around 4-5% of global Internet searches are related to healthcare information (Lin et al. 2016). While internet searches, online forums, and websites can help reduce the burden on medical professionals to provide information (Graf et al. 2022), lay users face a number of challenges in using these sources. These include information overload (Graf et al. 2022; Mohammed et al. 2020), outdated or incorrect information, especially for user-generated content (Teplinsky et al. 2022), and low readability of medical text (Krishna et al. 2021; Jin et al. 2022).

These challenges can be tackled by jointly leveraging knowledge graphs (KGs) and large language models (LLMs) to develop question answer (QA) systems. KGs are a means of representing knowledge by concepts and their relationships through a graphical structure of nodes and edges (Ji et al. 2021). LMs are probability-based models that learn statistical properties of the sequential distribution of words in documents (Bengio 2008). While KGs can capture facts from textual data, LLMs excel at providing natural language answers to users' questions. Using KG and LLM techniques jointly to build automated healthcare question-answer (QA) systems can help resolve the challenges in answering lay users' healthcare questions. These QA systems are automated software systems that use natural language processing (NLP) to extract relevant information from a structured knowledge base (KB) (Zhu et al. 2021). The aim of these systems is to answer lay users' questions using easy to understand sentences devoid of complex medical jargon. These systems need to be thoroughly tested to make sure quality answers are furnished to lay users. This facilitates human-AI collaboration for users' healthcare needs.

There are two main forms of providing answers to users' questions in a natural QA setting: (1) extractive QA (Chen et al. 2017) which assumes that given a question, the answer is readily available in one of the many documents in-line in the KB, and (2) generative QA (Lewis et al 2020) which makes no assumption about the answer being directly present in a given KB. Generative QA infers and constructs an answer based on the understanding of the question and the potential matching sections (or text portions) of a given KB. Since it does not have restrictive assumptions, we make use of a generative approach.

This article first proposes evaluation dimensions for assessing the answers provided by a generative QA system. Next, it describes the results from evaluation of DiseaseGuru, a KG and generative LLM-based QA system we built for chronic disease QA. We benchmarked our system with respect to three other algorithms across these dimensions. Finally, we present metrics for lay users and medical professionals for a future field study to evaluate the QA system.

Background and Dimensions

Generative QA entails that the automated system go through a large number of documents in a KB to find the right sub-set of documents (called the contexts) and generate an answer from the KB documents. While generating answers from a large number of documents helps solve the information overload issue for end users, it is imperative to evaluate the quality of the generated answer (He et al. 2022).

For evaluation, the QA system-generated answers are typically compared against pre-provided answers also known as gold answers (Risch et al. 2021) to see how similar they are.

Earlier studies on healthcare QA systems restricted themselves to either lexical or semantic similarity (Esteva et al. 2021; Fecho et al. 2021; Graf et al. 2022) as dimensions to assess quality of QA system-generated answers with respect to gold answers. However, we propose that the answers should also be assessed for contradictions and readability as these can pose threats to the utility and acceptance of a QA system by lay users. Hence, we provide a more comprehensive set of four dimensions for evaluating healthcare QA systems i.e., lexical similarity, semantic similarity, readability, and contradictions. Next we discuss each of these individual dimension.

Lexical Similarity

Lexical based similarity metrics rely on string comparison (He et al. 2022) popularly known as n-gram comparison. In this category, IBM BLEU (BiLingual Evaluation Understudy) (Papineni et al., 2002) is among the first popular metrics proposed to evaluate the quality of machine translation (MT). Subsequently, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003) was proposed as an improvement over BLEU as an automated summary evaluation using n-grams (e.g., uni-gram, bi-gram) overlap. ROUGE based ranking of text summaries correlated well with human ranking. Another metric, METEOR (Banerjee and Lavie, 2005), was proposed as a fully automated word-to-word generalized uni-gram matching-based improvement on BLEU that filled the gap of absence of unigram-recall in BLEU. METEOR has only been evaluated for machine translation from English to Arabic and Chinese. Further, EM (Exact Match) and F1 score (Rajpurkar et al., 2016) were introduced for measuring the quality of machine generated short answers in the context of extractive QA. While EM measures the exact match between a gold answer and a system generated answer, the F1 score computes the average overlap between the gold answer and system generated answer.

For healthcare QA evaluation, we need to compare the word/phrase level overlap between answers from the healthcare QA system with the gold answers. Hence, the ROUGE metric is appropriate to measure lexical similarity because it is specifically designed to measure the overlap between two passages of text in the form of overlapping n-gram or sequences of words. ROUGE ranks a given text on the scale of 0 to 1 (lowest to highest) with respect to the text of a gold answer (or reference text). Also, ROUGE has outperformed other content-based evaluation metrics and is the most popular metric for lexical similarity (Yu 2022). Thus it is utilized in our study.

Semantic Similarity

Lexical similarity metrics have a shortcoming whereby these metrics rely on word-level overlaps but do not consider the semantic overlap (texts with similar meanings, synonymous words) between a system generated answer and a gold answer (Krishna et al. 2021; Risch et al. 2021; Jin et al. 2022). To address this issue, Semantic Text Similarity (STS) is useful, which measures the semantic overlap between a pair of text snippets (Reimers and Gurevych 2019).

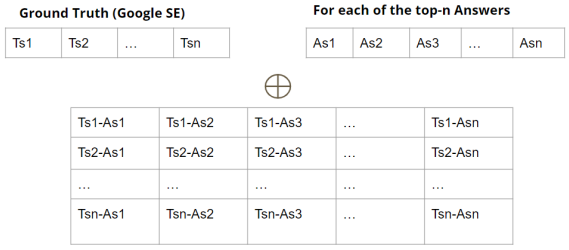

Figure 1: Pair-wise sentence entailment scoring. Tsn: nth sentence of the ground truth, Asn: nth sentence of the machine generated answer. A pairwise set is prepared and checked for any negative entailment between ground truth and system generated answer

In this regard, SemEval STS (Agirre et al. 2012) has promoted development of methods to measure the semantic similarity between a pair of texts. This includes MedSTS (Wang et al. 2020), which is a sentence similarity metric developed specifically for texts from the medical domain. However, since texts from the medical domain contain major variations in terms of synonymous words (Jin et al. 2022), it is also important to use a word-level semantic scoring metric such as BERTScore (Zhang et al. 2019). BERTScore has been shown to correlate well with human-judgements on sentence-level similarity. Hence we propose its use for healthcare QA system evaluation. BERTScore measures the similarity between word embeddings using cosine similarity (scale of 0 to 1; lowest to highest) thereby recognizing words which are synonymous to each other.

Readability and Contradictions

Lexical and semantic similarity metrics are unable to assess the readability dimension of a system generated text. Yet, low readability of medical content is an issue for lay readers (Graf et al. 2022). Also, reading and understanding a long paragraph of text becomes challenging for such users (Krishna et al. 2021). Hence, for lay users, there is a need to produce relatively short and precise answers. One of the most popular readability metrics is the Flesch-Kincaid score (FK-score; Kincaid et al. 1975) which correlates highly (r = 0.91) with measures of reading tests. Flesch-Kincaid score uses features such as total number of words, number of sentences, and total syllables in a given text. A score for a given text is generated on a scale of 0 to 100 (lowest to highest) with higher scores signaling easier to read text. We propose FK-score as a metric to assess the readability dimension for QA system evaluation.

Additionally, language models are prone to hallucinations (factual inconsistency) in their responses (Ji et al. 2023). Therefore, in addition to readability, we propose that researchers should also evaluate the dimension of contradictions. This is to test for any contradictory or inconsistent sentences in the system generated answers which can mislead lay users. To this end, the approach of sentence entailment can be used for QA system evaluation. Two sentence vectors in natural language will point in the same direction if they convey that same information and will point in op-

	Benchmark-1 (ChatGPT)	**Benchmark-2** (Graf et al. 2022)	**Benchmark-3** (Esteva et al. 2021)	**DiseaseGuru** (LFQA joint-reasoning)
ROUGE (mean)	0.20 ± 0.1 (6.4e-10)$^{+}$	0.23 ± 0.3 (2.0e-05)$^{+}$	0.18 ± 0.1 (1.8e-11)$^{+}$	**0.40** ± 0.2
MedSTS (mean)	2.53 ± 0.7 (7.4e-06)$^{+}$	1.9 ± 1.2 (8.0e-12)$^{+}$	2.33 ± 0.6 (2.9e-10)$^{+}$	**3.03** ± 0.1
BERTScore (mean)	0.86 ± 0.0 (2.6e-13)$^{+}$	0.84 ± 0.1 (3.3e-10)$^{+}$	0.85 ± 0.0 (5.0e-15)$^{+}$	**0.90** ± 0.0
FK-Score (mean)	55.47 ± 12.9 (0.37)$^{+}$	53.5 ± 37.2 (0.02)$^{+}$	62.30 ± 26.5 (0.04)$^{+}$	**64.06** ± 16.1 (0.02)
Contradict-ions (mean)	0	3	3.5	**0**

Table 1: Benchmarking DiseaseGuru

posite directions if they provide contradictory information (Martín et al. 2022). The sentence entailment approach generates mainly three outputs i.e., positive, neutral, and negative entailment between a given pair of texts. To uncover contradictory information, a pair-wise entailment computation of each answer sentence in the system generated text with respect to each sentence in the gold answer text can be performed (Figure 1).

Case of DiseaseGuru

DiseaseGuru[1] is a generative QA system we built using a multi-component approach. It consists of three subcomponents: (1) a data base of text documents stored in vector format (VDB), (2) a fine-tuned sentence transformer, and (3) a disease KG. The VDB plays the role of our system's KB with all the text documents stored in a vector (or numeric) format. The sentence transformer is enriched for domain knowledge using domain adaptation and helps generate a set of potential answers from LLM trained on Long-Form Question Answering (LFQA) (Blagojevic 2022). Finally, disease KG provides a mechanism to select the best answer from the potential answers of the LLM.

We evaluated DiseaseGuru against multiple benchmarks. Our first benchmark-1 is the state-of-the-art chatbot system ChatGPT-3.5 from OpenAI[2]. Our second benchmark-2 was constructed to be similar to COVID-19 QA systems described in recent research which used extractive QA techniques to answer healthcare questions (Graf et al. 2022). Finally, our third benchmark-3, was constructed to be similar to a generative QA system proposed in the healthcare context (Esteva et al. 2021). We performed a rigorous evaluation of the answers using the metrics under the proposed dimensions of lexical similarity, semantic similarity, readability, and contradictions.

We evaluated DiseaseGuru for the top 10 most common

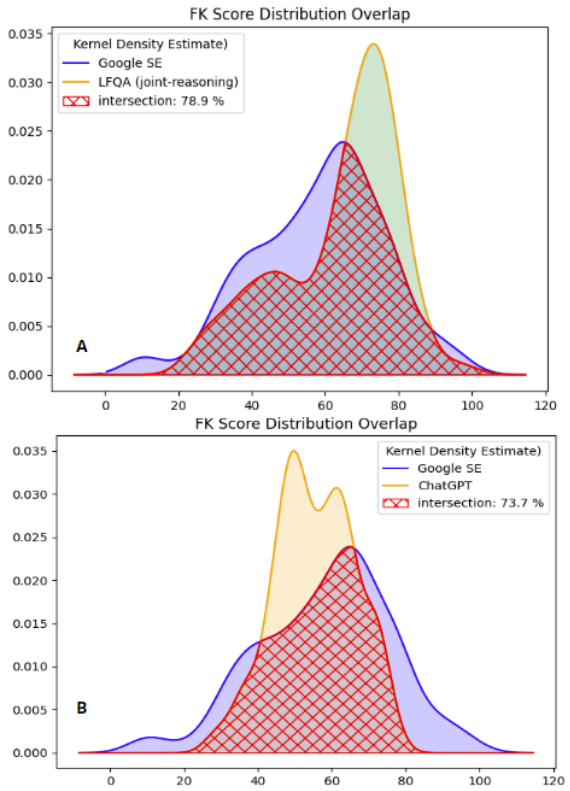

Figure 2: Kernel Density Distribution of Gold-Ans (blue), 2A. DiseaseGuru (in green) and 2B. ChatGPT (in orange)

chronic diseases[3]. These diseases were - Alzheimer's disease, Asthma, Chronic Cough, Chronic Kidney Disease, Cancer (Colorectal Cancer, Leukemia), Diabetes, Cardiac issues (Heart Attack, Heart Failure), Obesity, Osteoarthritis, and Periodontitis. In total we obtained a set of 75 test questions from MedQuAD (Ben Abacha and Demner-Fushman 2019). For each question, a gold answer was generated using arguably the most popular search engine i.e., Google Search engine (SE) (Urman et al. 2022).

Results

We performed a rigorous answer quality evaluation using metrics which evaluate lexical similarity, semantic similarity, readability, and contradictions (see Table 1). On the popular lexical metric ROUGE, our system significantly outperformed ChatGPT-3.5 and the other two benchmarks with a mean score of 0.4. On semantic metrics, MedSTS and BERTScore, our system significantly outperformed the benchmarks with scores of 3.03 and 0.9 respectively. On the FK-score for readability, our system showed higher readability in generated answers compared to the benchmark systems. Finally, our system did not generate contradictory information when compared to the gold answers.

Additionally, the distributional overlap between the systems was inspected using kernel density estimate plots (see

[1]Details of DiseaseGuru are in another publication – undisclosed due to anonymity requirements

[2]https://openai.com/blog/chatgpt/

[3]https://www.cdc.gov/chronicdisease/about/costs/index.htm

Figure 2). The percentage overlap shows that the readability of our system's answers (see Figure 2A) (78.9%) are better than that of Google SE as the distribution, including the mode, is shifted to the right. In comparison, ChatGPT-3.5 answers achieved a lower overlap (73.7%) with those from Google SE (see Figure 2B).

Contributions and Future Work

For QA systems developed for lay users, we propose that it is imperative to assess the quality of system responses with respect to multiple dimensions covering lexical similarity, semantic similarity, readability, and contradictions. In this paper, we contribute by proposing a comprehensive set of evaluation dimensions for QA systems for lay users. We then use these dimensions to evaluate DiseaseGuru, a healthcare QA system we developed that integrates KG and LLM techniques to answer common questions related to chronic diseases. The results show that the system outperformed ChatGPT-3.5 and two other benchmark systems for disease QA. A holistic evaluation can increase the trust in the system's utility. It can thereby promote system adoption, use and healthcare collaboration outcomes for lay users. In future, we will design user evaluation scales for both lay users and medical professionals. For lay users, we will measure their perceptions such as relevance and comprehensibility of answers from the QA system. For medical practitioners, we will assess their perceptions like completeness, informativeness, and accuracy of answers (adapted from Zhu et al. 2009) from the QA system. These perceptions will be assessed through field studies of the QA system with lay users and medical professionals.

Acknowledgements

This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016).

References

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 385–393.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Ben Abacha, A.; and Demner-Fushman, D. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1): 1–23.

Bengio, Y. 2008. Neural net language models. *Scholarpedia*, 3(1): 3881.

Blagojevic, V. 2023. Long-Form QA beyond ELI5: an updated dataset and approach. https://towardsdatascience.com/long-form-qa-beyond-eli5-an-updated-dataset-and-approach-319cb841aabb. Accessed: 2023-03-06.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Esteva, A.; Kale, A.; Paulus, R.; Hashimoto, K.; Yin, W.; Radev, D.; and Socher, R. 2021. COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1): 68.

Fecho, K.; Bizon, C.; Miller, F.; Schurman, S.; Schmitt, C.; Xue, W.; Morton, K.; Wang, P.; Tropsha, A.; et al. 2021. A biomedical knowledge graph system to propose mechanistic hypotheses for real-world environmental health observations: cohort study and informatics application. *JMIR Medical Informatics*, 9(7): e26714.

Graf, J.; Lancho, G.; Zschech, P.; and Heinrich, K. 2022. Where Was COVID-19 First Discovered? Designing a Question-Answering System for Pandemic Situations. *arXiv preprint arXiv:2204.08787*.

He, J.-W.; Jiang, W.-J.; Chen, G.-B.; Le, Y.-Q.; and Ding, X.-F. 2022. Enhancing N-Gram Based Metrics with Semantics for Better Evaluation of Abstractive Text Summarization. *Journal of Computer Science and Technology*, 37(5): 1118–1133.

Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; and Philip, S. Y. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2): 494–514.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Jin, Q.; Yuan, Z.; Xiong, G.; Yu, Q.; Ying, H.; Tan, C.; Chen, M.; Huang, S.; Liu, X.; and Yu, S. 2022. Biomedical question answering: a survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2): 1–36.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Krishna, K.; Roy, A.; and Iyyer, M. 2021. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Lin, C.-Y.; and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, 150–157.

Lin, Y.-L.; Chung, C.-Y.; Kuo, C.-W.; and Chang, T.-M. 2016. Modeling health care Q&A questions with ensemble classification approaches. In *AMCIS 2016 Proceedings. 6.*

Martín, A.; Huertas-Tato, J.; Huertas-García, Á.; Villar-Rodríguez, G.; and Camacho, D. 2022. FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-Based Systems*, 251: 109265.

Mohamed, A.; Parambath, S.; Kaoudi, Z.; and Aboulnaga, A. 2020. Popularity agnostic evaluation of knowledge graph embeddings. In *Conference on Uncertainty in Artificial Intelligence*, 1059–1068. PMLR.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Risch, J.; Möller, T.; Gutsch, J.; and Pietsch, M. 2021. Semantic answer similarity for evaluating question answering models. *arXiv preprint arXiv:2108.06130*.

Saggion, H.; Radev, D.; Teufel, S.; and Lam, W. 2002. Meta-evaluation of summaries in a cross-lingual environment using content-based metrics. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Teplinsky, E.; Ponce, S. B.; Drake, E. K.; Garcia, A. M.; Loeb, S.; van Londen, G.; Teoh, D.; Thompson, M.; Schapira, L.; and for Outcomes using Social Media in Oncology (COSMO), C. 2022. Online medical misinformation in cancer: Distinguishing fact from fiction. *JCO oncology practice*, 18(8): 584–589.

Urman, A.; Makhortykh, M.; and Ulloa, R. 2022. The matter of chance: auditing web search results related to the 2020 US presidential primary elections across six search engines. *Social science computer review*, 40(5): 1323–1339.

Wang, Y.; Afzal, N.; Fu, S.; Wang, L.; Shen, F.; Rastegar-Mojarad, M.; and Liu, H. 2020. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54: 57–72.

Yu, H. 2022. Survey of Query-based Text Summarization. *arXiv preprint arXiv:2211.11548*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhu, F.; Lei, W.; Wang, C.; Zheng, J.; Poria, S.; and Chua, T.-S. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

Zhu, Z.; Bernhard, D.; and Gurevych, I. 2009. A multi-dimensional model for assessing the quality of answers in social Q&A sites. In *ICIQ*, 264–265.