Question Formulation and Transformer Model Resilience*

*Effects of question-wording variants on the robustness of transformer-based question-answer models

Ashraf Elnashar
Dept. of Computer Science
Vanderbilt University
Nashville, Tennessee, USA
ashraf.elnashar@vanderbilt.edu

William Schreiber
Dept. of Computer Science
Vanderbilt University
Nashville, Tennessee, USA
william.schreiber@vanderbilt.edu

Jules White
Dept. of Computer Science
Vanderbilt University
Nashville, Tennessee, USA
jules.white@vanderbilt.edu

Douglas C. Schmidt
Dept. of Computer Science
Vanderbilt University
Nashville, Tennessee, USA
d.schmidt@vanderbilt.edu

***Abstract*—Question-answer is a paradigm that seeks to provide automated responses to queries posed in natural language utilizing a body of textual content as the source of the answers. A key research challenge is how the changes in question formulation affect the stability of current question-answer transformer models. This paper conducts a preliminary analysis of the stability of question-answer transformer models in the medical domain when the same question is asked in different orders or with other semantically identical variations. The results from our experiments demonstrate that the arrangement of words influences the outcome and consistency of answers from transformer models.**

***Index Terms*—NLP, trasnformer model, question-answer model, question formulation, COVID-19**

I. Introduction

The application of natural language processing (NLP) to automate question answering has significant potential in the field of medicine [gCCEY10]. There has been growing interest in deriving answers to healthcare questions from the unstructured data included in medical literature [LR16]. Question answering systems enable physicians to ask questions and receive answers in natural language [gCCEY10].

For example, rather than wasting valuable time manually searching for the answer to the question from a set of complex research papers, doctors can employ a question answering model to extract critical data that is related to the question. This approach not only allows the physician to focus on presenting the answer to a patient, but also ensures the physician is more confident in the answer [gCCEY10].

Deep learning has increased the performance of question answering models. Transformer models are best performing deep learning approach for question answering today [VSP$^+$17]. These models are based on the concept of "attention," which enables the model to concentrate on a portion of the sequence while simultaneously predicting the response.

The key innovation in transformers is that the overall ordering of the words from the question and the surrounding context of each word plays a significant role in formulating the response. In particular, transformers use differential weighting to indicate which other parts in the sentence are most important to the understanding of a particular word. For example, in the question *"What is the size (diameter) of the SARS-CoV-2 virus?"* the terms "size," "(diameter)," "SARS-CoV-2," and "virus" are the most essential to comprehend.

A question reordering approach can help to resolve a number of uncertainties about the meaning of specific words. In the context of natural language processing, transformers were first established to eliminate recursion and allow parallel computing (to save training time). Transformers also help to reduce performance degradation due to long dependencies.

As outlined above, transformers are highly context-dependent in how they interpret questions. The individual understanding of words is based on a weighting of the surrounding words. A key research challenge, therefore, is determining whether transformers can consistently provide correct answers to questions, regardless of the ordering of the words in the questions.

For example, two questions may be phrased with the same words, but have different word orderings with the same meaning, such as "What is the size (diameter) of the COVID virus?" vs. "What is the COVID virus (diameter) size?" Although these two questions have slightly different wordings they are essentially the same inquiry. Ideally, a transformer should produce the same answer for each word ordering.

This paper uses a body of medical literature and a large-scale dataset of questions regarding the coronavirus disease 2019 (COVID-19) to investigate the impact of word ordering on the answers produced by transformer-based question answering models. Our work composes semantically identical questions using the same words—but with different word orderings—and then measures answer stability. We investigate whether or not the transformer model can produce consistent answers or if additional research is needed on architectural components of these models to improve stability with respect to word ordering. In addition, we measure the overall accuracy of question answering models on our large-scale dataset.

The remainder of this paper is organized as follows: Section II gives an overview of our datasets and transformer model, Section III analyzes results from experiments we conducted, and Section IV presents concluding remarks and outlines future work.

II. Overview of Our Datasets and Transformer Model

1) Datasets: We used the COVID-19 question-answer community dataset obtained from biology stack exchange website, which includes 642 COVID-19 questions answers. Likewise, the medical literature CORD-19 *”COVID-19 Open Research Dataset”* included in [WLC^{+}20] were initially sourced from PubMed Central (PMC), PubMed, the Covid-19 Database.

2) Transformer model: Our work in this paper is based upon the deepset/roberta-base-squad2-covid transformer model. This model applied the ”Bidirectional Encoder Representations from Transformers” (BERT) technique. This model was fine-tuned with a question answer approach and trained on CORD-19 SQuAD-style annotations outlined above.

To improve semantic similarity with relation to word reordering, our transformer model adds a positional encoding vector to the encoder’s input, which comprehends the order of words in an input sequence or inquiry, as show in Figure 1. This positional embedding vector provides the transformer with information regarding the positions of the input vectors. Instead of being concatenated, the positional vectors are simply appended to the relevant input vectors. The positional embedding of the input to the encoder can be improved by boosting the word order position information, which in turn leads to increased semantic similarity.

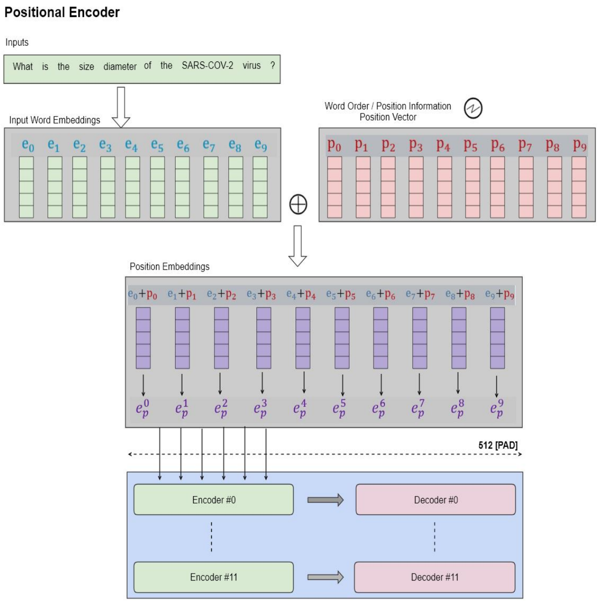
Positional Encoder
Inputs
What is the size diameter of the SARS-COV-2 virus ?
Input Word Embeddings
e_0 e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8 e_9
Word Order / Position Information
Position Vector
p_0 p_1 p_2 p_3 p_4 p_5 p_6 p_7 p_8 p_9
⊕
Position Embeddings
e_0+p_0 e_1+p_1 e_2+p_2 e_3+p_3 e_4+p_4 e_5+p_5 e_6+p_6 e_7+p_7 e_8+p_8 e_9+p_9
e_p^0 e_p^1 e_p^2 e_p^3 e_p^4 e_p^5 e_p^6 e_p^7 e_p^8 e_p^9
512 [PAD]
Encoder #0
Decoder #0
Encoder #11
Decoder #11

Fig. 1. Embedding Position Encoding

III. Analysis of Experimental Results

We conducted experiments to determine if the question-answer transformer model delivered consistent answers when the words in questions were rearranged to produce inquiries with the same semantics and expected answer. To perform these experiments, we used a COVID-19 dataset to feed questions generated by users and relevant medical literature to the transformer model, which then generated answers. We then assessed the stability of the answers produced by the model by comparing the answers to the original and modified questions, as described below.

A. Impact of a question’s random keyword reordering on the consistency of the answers

This first experiment examined how rearranging words in questions, while retaining the question’s original semantics, impacts the consistency of the answers. In particular, we evaluate whether the model produces the same answer to the question consistently, regardless of word reordering.

The experiment initially captured baseline answers to the original question word ordering that were subsequently used to compare against answers from the modified question word orderings. The specific transformer models we experimented with were fine tuned with question-answer approaches and trained using the COVID-19 dataset. We fed each question from the question set through each model and generated the answers. We also compared the answers produced by the model to the ground truth answer in the dataset using the BiLingual Evaluation Understudy (BLEU) score [PRWZ02], which employ n-gram co-occurrence statistics ranging from 0 to 1 to measure the similarity of machine-translated text to a set of high quality reference translations. Figure 2 shows the overall question answering accuracy of the baseline models.

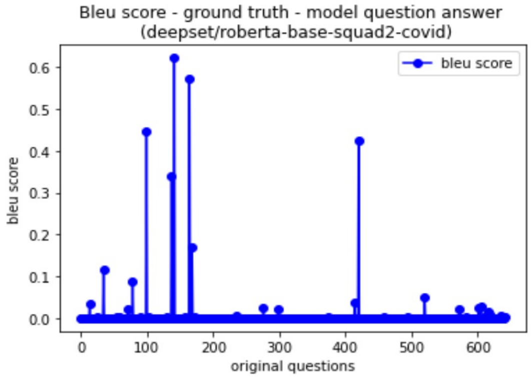

Fig. 2. BLEU Score Comparing Model and Ground Truth Answers.

We then produced reworded questions that preserved the semantics of the overall question by collecting keywords from the questions and rearranging them randomly in the questions. Keywords in the questions were chosen using Keyword Extraction Algorithms [TMM19] and rearranged randomly. The key variable in this experiment was “choose one word for one position move.” Rearranging keywords in questions while retaining question semantics was the most challenging aspect of the experiment.

Although we preserved each question’s semantics , we needed to measure how different the original word reordering was vs. the baseline question. Various metrics measure the distance between strings, ranging from low-level metrics (such as hamming distance [NFS12]) to character-level metrics (such as "edit distance"). Our experiments used word-level edit distance, which determines the number of words that must be swapped to reach the original sentence.

The edit distance d_{mn} between two terms a and b was defined by [HM11] to assume that the original question "(III-A)"contains the terms $a = a_1 \dots a_m$, and that the amended question contains the terms $b = b_1 \dots b_n$, and that the edit distance d_{mn} is the least weight sequence w of change operations that transforms a to b. These calculations measure the distance between two strings by converting an input word (a) into an output word (b), excluding the substitution of a character, which has no cost. Positive weight functions are represented by the operations $w_{ins}(x)$,$w_{del}(x)$, and $w_{sub}(x, y)$, as shown below:

$$\mathrm{d}_{i0} = \textstyle\sum_{k=1}^{i} w_{del}(a_k) \qquad \text{for } 1 \leq i \leq m$$
$$d_{0j} = \textstyle\sum_{k=1}^{j} w_{ins}(b_k) \qquad \text{for } 1 \leq j \leq n$$
$$d_{ij} = \begin{cases} \mathrm{d}_{i-1,j-1} \text{ for } a_i = b_j \\ \text{min=}\begin{cases} \mathrm{d}_{i-1,j} + w_{del}(a_i) & for\ \mathrm{a}_i \neq b_j \\ \mathrm{d}_{i,j-1} + w_{ins}(b_j) & 1 \leq i \leq m \\ \mathrm{d}_{i-1,j-1} + w_{sub}(a_i, b_j) & 1 \leq j \leq n \end{cases} \end{cases}$$
(1)

The edit distance between each modified question and the original question was based on the number of word swaps, rather than at the character level. The BLEU score can be calculated using both the answers to the original questions and the answers to the modified questions. By graphing the distance against the BLEU score, we determined the impact that the distance has on the score.

The experiment steps using this word reordering approach are as follows: (1) Utilize Rapid Keyword Extraction (RAKE) to extract pertinent keywords from each of the 642 questions in the original community dataset and choose the two words that most accurately express the question's meaning, (2) swap the keywords' places, (3) calculate the word edit distance between the original question and the revised version [RY98], (4) use each transformer model to produce an answer to the reordered question, and (5) compare the original answer to the new answer using the BLEU score. A BLEU score of 1.0 indicated that identical answers were produced for the original answer, while decreasing BLEU scores indicate the answers diverged further from the original answer, indicating a loss of consistency in the answers provided.

The formula "(III-A)" was used to compute and construct the total number of modified questions from the original questions, as shown below.

$$mq_k = \textstyle\sum_{i=1}^{k} (k_i! \ * \ q_i) \qquad \text{for } 1 \leq i \leq k$$
where mq = "# of modified questions"
and k = "extracted keywords"
and q = "# of questions".
$$oq_k = \textstyle\sum_{i=1}^{k} (q_i) \qquad \text{for } 1 \leq i \leq k$$
where oq = "# of original questions"
and q = "# of questions".
(2)

Formula "(III-A)" shows how the experiment used 318 of the original questions and produced 91,074 modified questions, which was sufficient for this experiment. We only used 318 of the original questions since the maximum distance is 6. If the distance was changed to 7, however, the total number of original questions would increase to 429, which would yield more than 650,000 modified questions. We capped the experiment at an edit distance of 6 due to the computational cost of processing an additional 0.5 million questions.

Our question-answer transformer model next answered 91,074 modified questions from the original 318 questions. The original questions' answers (which considered the BLEU reference corpus) and the modified questions' answers (which considered the BLEU hypotheses) were then used to compute the BLEU score and validate the answer's stability. Figure 3 depicts the average BLEU at each distance, which decreased as distance increased. This result indicates that the transformer model's answer was unstable when the order of the keywords in the original question changed.

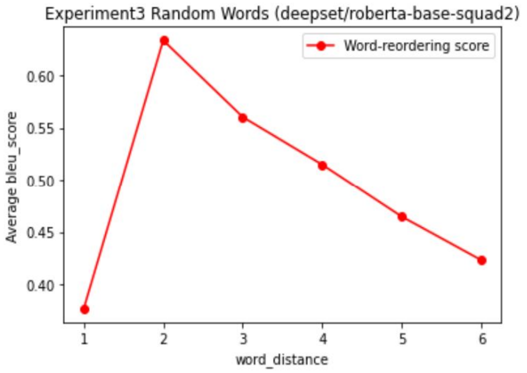

Fig. 3. Word Distance vs. Average BLEU Score

B. Semantic similarity between the original question and the modified question

This second experiment determined the degree to which the changed question and the original question shared semantic content. The similarity between the original and modified questions based on the distance indicated how reordering the original question would affect its semantics. Cosine similarity is a measurement commonly used to determine the score for semantic similarity.

Figure 4 depicts the outcomes of computing the average degree of semantic similarity as a function of distance. This figure shows how the average semantic similarity between the original and amended questions decreased as the word distance between them increased. The difference between the minimum and maximum average semantic similarity was 0.07. Despite a diminishing overall semantic similarity, the lowest semantic similarity was 0.92. This relatively high value indicated that the original and modified questions had similar meanings.

Demonstrating semantic similarity between the original and modified questions is important because the transformer model

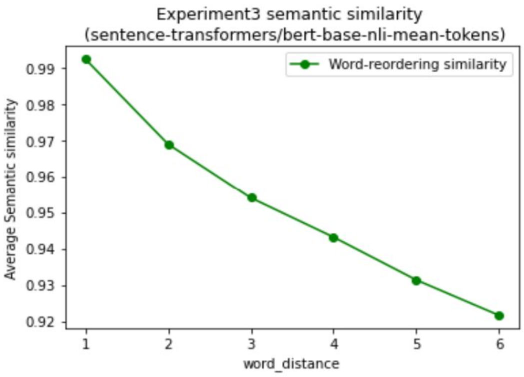

Fig. 4. Word Distance vs. Average Semantic Similarity

should deliver a consistent answer when the keywords of the original questions were rearranged if the semantic similarity was similar. It is also important to know the frequency of each semantic similarity score because this indicates which similarity score has the greatest number of questions.

The semantic similarity histogram in Figure 5 shows the majority of the amended questions had semantic similarities to the original questions with a score of 0.95. This figure indicates that most of the original and modified questions were semantically similar. Moreover, a few questions with similarity scores less than 0.8 can be disregarded because the percentage was so small compared to the 91,074 questions.

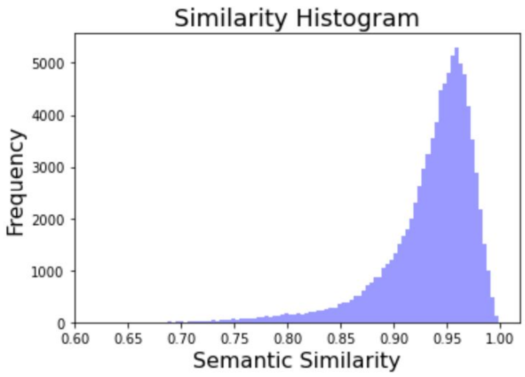

Fig. 5. Semantic Similarity Histogram

The results of this experiment showed that reordering keywords on the original question preserved the semantics of the original questions. This result was demonstrated by calculating the cosine similarity between the original and modified questions. Moreover, regardless of the distance between the original and changed questions, the semantic similarity were relatively close.

IV. Concluding Remarks

This paper analyzed the results of experiments that evaluated the degree to which the question-answer transformer model provides consistent answers when question keywords are rearranged in a manner that preserves semantic similarity. A COVID-19 dataset was used to feed questions and medical literature into the transformer model, which produced answers. The stability of the model was then evaluated by generating a BLEU score and comparing model output to the original and changed queries.

The following are a summary of lessons learned from our work reported in this paper:

- When the order of the keywords in the original question was changed, the output of a transformer model became unstable.
- The semantics of the original questions were preserved as a result of reordering keywords in the original questions, as demonstrated by cosine similarity between the original and changed questions. Rearranging terms in the original questions yielded the modified questions.
- It is critical to improve semantic similarity with relation to word reordering using the transformer model since higher semantic similarity yields a more stable answers to questions.

To address the instability of the transformer model our future work will study the encoding of the transformer's hidden layers to identify why the model loses the query's semantics during the encoding process.

References

[gCCEY10] Yong gang Cao, James J. Cimino, John Ely, and Hong Yu. Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43(6):962–971, 2010.

[HM11] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *ArXiv*, abs/1101.1232, 2011.

[LR16] Sweta P. Lende and M. M. Raghuwanshi. Question answering system on education acts using nlp techniques. In *2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, pages 1–6, 2016.

[NFS12] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[PRWZ02] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[RY98] Eric Ristad and Peter Yianilos. Learning string edit distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20:522 – 532, 06 1998.

[TMM19] M G Thushara, Tadi Mownika, and Ritika Mangamuru. A comparative study on different keyword extraction algorithms. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pages 969–973, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[WLC+20] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. Cord-19: The covid-19 open research dataset, 2020.