EXPLORING QUESTION ANSWERING AND GENERATION USING NEURAL MODELS AND SYNTACTIC TRANSFORMATIONS ON AN ENGLISH AND MULTILINGUAL CORPUS

Blessed Guda
Nebiyou Daniel Hailemariam
Oluwadara Adedeji
Chukwuemeka Malachi Ugwu

Abstract

This research explores the use of neural models for question answering (QA) and syntactic transformations and neural models for question generation (QG). For the QA system, the question type is first classified as either factoid or polar using Finetuned DistilBERT model trained on BookCorpus and English Wikipedia, which achieved an accuracy of 99.96%. For the factoid questions, the QA system uses a Finetuned BERT large model trained on BookCorpus and English Wikipedia and finetuned on the SQuAD dataset for question-answering. For polar questions, the context and question are passed to the Finetuned Roberta model to output a yes or no answer. The system achieves an f1 score of 91.3% and an Exact Match of 86.91% with the BERT large model for answering factoid questions. This research also explores the MT5 model on the MLQA dataset to answer questions from high-resource language (English) using content from low-resource languages. Furthermore, the research investigates the use of syntactic transformations and T5 neural models in the QG system. It is observed that the syntactic transformation system requires domain-specific linguistic knowledge to improve its performance. However, despite this limitation, the system can generate approximately 30 to 40% of the total generated questions fluently and answerably. To improve this, a T5 model trained on the SQuAD dataset was used for generating Wh-questions and a BoolQ dataset for polar questions. The manual evaluation shows that this question generation system produced over 95% of fluent and answerable questions from a pool of 20 questions generated.

1 Introduction

1.1 Question Answering

Question Answering (QA) is a subfield of Natural Language Processing (NLP) that aims to build systems that can automatically answer questions posed by humans in natural language (Ostapov, 2011). QA systems can be classified into two categories: open-domain and closed-domain. Open-domain QA systems aim to answer any question that a user might ask, while closed-domain QA systems are designed to answer questions within a specific domain (Lende and Raghuwanshi, 2016).

Deep learning methods have been used extensively in recent years for QA tasks due to their ability to learn complex patterns in data and their ability to generalize well on unseen data (Sarker, 2021). Deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers have been used for QA tasks (Chaturvedi, 2018)(Hao et al., 2022)(Sarker, 2021). However, deep learning methods require large amounts of labeled data and computational resources (Hao et al., 2022)(Sarker, 2021).

Traditional methods such as Information Retrieval (IR) and Rule-based methods have also been used for QA tasks (Adnan and Akbar, 2019). IR-based methods retrieve relevant documents from a large corpus of text based on keyword matching or semantic similarity (Zhou et al., 2020). Rule-based methods use hand-crafted rules to extract answers from text (Rivas et al., 2019). Traditional methods are less computationally expensive than deep learning methods and require less labeled data (Adnan and Akbar, 2019). However, traditional methods may not perform well on complex questions or questions that require reasoning (Zhou et al., 2020).

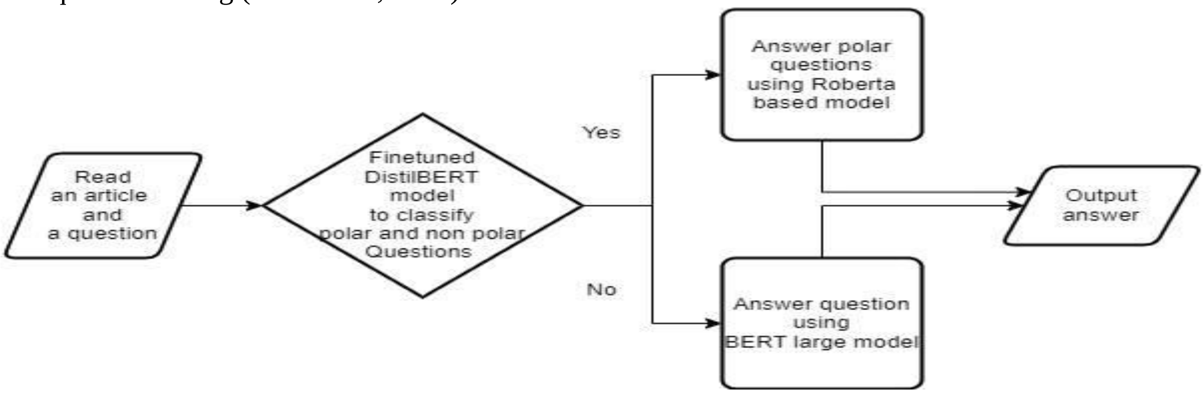

Figure 1: Architecture of Question Answering System

1.2 Question Generation

Question Generation (QG) is a field of study in natural language processing that focuses on developing systems that generate questions automatically using different sources such as unprocessed text, semantic representation, or databases (Pan et al., 2019). It is useful for educational purposes (Le et al., 2014) or to generate questions and answers for training and improving a Question Answering system (Mulla and Gharpure, 2023) and reduce the time needed for human labor to annotate question-answer datasets (Zhang et al., 2022). Different approaches have been explored to generate questions. These approaches to question generation can be template-based (Fabbri et al., 2020), syntactic transformations based or neural model-based (Mulla and Gharpure, 2023).

In template-based matching, a sentence similar to a context being considered is retrieved from the corpus, and a question is generated from this sentence (Fabbri et al., 2020). In syntactic transformation, sentences from a passage are transformed based on the constituency and dependency parse of the sentence (the syntax of the sentence) (Varga and Ha, 2010). Transformation rules are then used to generate questions based on the on the syntax of the sentence.

The other approach to question generation is the neural model approach which uses the encoder-decoder architecture to generate questions (Dwivedi and Marappan, 2009). This encoder-decoder model is trained on datasets such as SQuAD, which contain passages, questions and questions. Typically, the questions are concatenated with the corresponding passages and fed into the decoder. The encoding representation of the question is used by the decoder in generating the answer. Some works have also extracted more features such as Style and Clue (Liu et al., 2020) to make the questions generated to be humanlike. This neural-based task of generating questions can be performed either from scratch or by fine-tuning previously pre-trained models. This is the contemporary way of question generation as it generates better questions which are harder to answer and not limited to the syntax sentences as in syntactic transformation (Du et al., 2017; however, they require a huge amount of training data and computational resources (Zohuri, 2020).

2 System Architecture

2.1 Question Answering

Figure 1 depicts the architecture of the complete QA system. After reading the context and questions, each question is passed to the Finetuned DistilBERT model to classify if the question is polar or non-polar. The Finetuned DistilBERT model is a transformer model trained with BERT serving as a teacher to guide its training on BookCorpus and English Wikipedia, the same datasets the BERT base model was trained on (Sanh et al., 2019). It was trained on the None dataset (Question Classifier V2 (2023)) and has achieved an accuracy of 99.96%. If the question is classified as a polar question, it is passed to the Finetuned Roberta model - a model that enhances the BERT model using larger mini-batches and learning rates during training. This subsystem outputs a yes or no answer to the polar question.

On the other hand, if the Finetuned DistilBERT model classifies the question as non-polar, the question is passed to the BERT large model, which is pre-trained on BookCorpus and English Wikipedia and finetuned on the SQuAD dataset for question-answering purposes (Devlin et al., 2019). The BERT large model is trained to read a given passage of text and provide an answer to a question along with its probability. The system applies a threshold score greater than 0.3 to the probability score to determine if an answer is correct. If the probability score is greater than or equal to 0.3, the system returns the answer as the final output. If the probability score is lower than 0.3, it returns an empty string as the answer. Returning an empty string means that the question is unanswerable given the input passage.

This research also explores multilingual QA systems. The research in (Asai et al., 2021) collected a multilingual dataset (XOR-TYDI QA) of questions from 7 languages. The questions were collected from native speakers of those languages rather than using other existing approaches that produced questions through translation. This research investigates multilingual QA systems. The motivation for this is that when questions are asked of information not directly related to the culture of the speaker, they are less likely to have answers in the questioner's language (Information asymmetry) (Ewa S. Callahan, 2013). The approach in (Asai et al., 2021) used content from a high-resource language (English) to answer questions from a low-resource language.

However, research was not conducted to answer questions from a high-resource language using content from low-resource languages. This approach may help in answering questions in a language with high resources, such as English, using resources from low-resource languages. It may be helpful for questions with cultural inclinations tied to the low-resource language that may not be captured by the resources in English. To investigate this, the research explores multilingual QA systems using the MT5 model and the MLQA dataset that contains English-to-Vietnamese, English-to-Arabic, and English-to-Spanish translation datasets. The aim is to answer questions from a high-resource language (English) using content from low-resource languages.

2.2 Question generation

The methodology explored in this project for question generation is syntactic transformations and neural models. In this section, we explain each methodology explored.

2.2.1 Syntactic Transformation

Figure 2 depicts the system architecture for the syntactic transformation process. The first step involves preprocessing the document, during which all pronouns are replaced with their references through coreference resolution. It is done using the Coreferee module within the spaCy Python library. Subsequently, the subject, verbs, and objects were extracted using the dependency tree technique in spaCy. It involved looping through all the tokens to obtain the root token. The root token is the verb on which other tokens depend.

After obtaining the root token, we iterate through all its children. Two lists of probable subject and object dependencies are used as a lookup for identifying if children of the root are subjects or objects. We loop through all the children of the root token. If the children are in the predefined subject dependencies list, they are marked as subjects. Likewise, if the children are in the predefined objects dependencies list, they are marked as objects. Finally, the subject, root token (verb), and object are returned.

The subjects and objects that are returned also have dependent tokens, which provide additional meaning to them. The subtrees are extracted and form the subject and object words or phrases. Subsequently, transformations were performed using the Named Entity Recognition (NER) tags. For example, if the NER tag is DATE or TIME, the

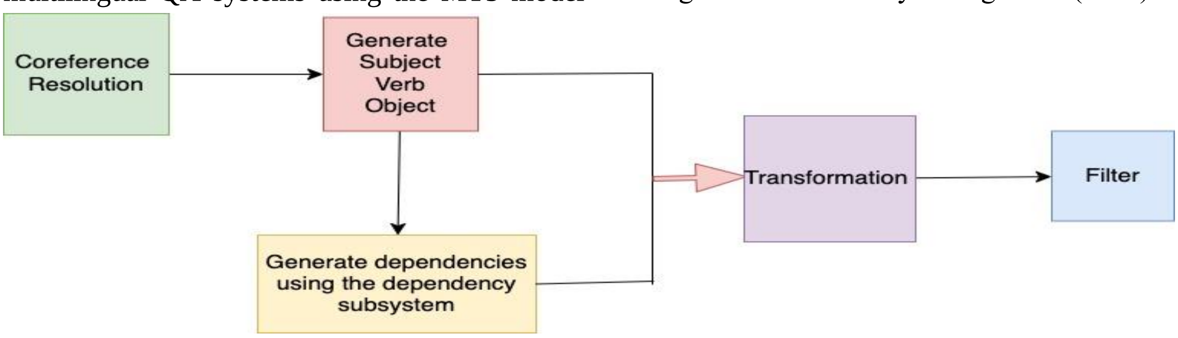

Figure 2: Syntactic transformation architecture.

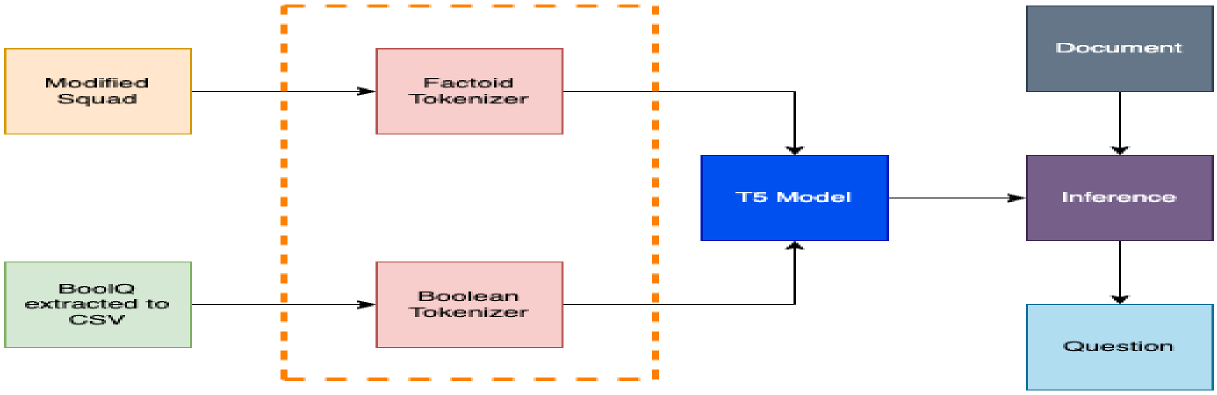

Figure 3: System diagram for the Neural model

question transformation will be "When did" + subject text + verb inflection. The verb inflection was applied. This was necessary to make the questions grammatical and fluent. However, the questions generated from the syntactic transformation were not good enough. A filtering technique was explored. However, due to the simplicity of the rules implemented the quality of the questions generated was low. Hence, a neural network approach was explored.

2.2.2 Neural network model

Figure 3 shows the system architecture for the neural model. The Text-to-Text Transfer Transformer (T5) is a model based on the encoder-decoder architecture (Raffel et al., 2020) and is suitable for text-to-text problems such as question generation. The T5 model was trained twice, one for generating Wh-questions and the other for polar questions. For the Wh-questions, the model was trained using the SQuAD dataset, while the polar QA were trained on the BoolQ dataset.

Each input question-passage pair was concatenated, tokenized, and terminated with an end-of-sequence token. Both models were trained and evaluated on the train and validation sets defined in the original datasets without any modifications. The factoid T5 model was trained for 5 epochs with a learning rate of 0.0001. Similarly, the polar T5 model was trained for 4 epochs, using a learning rate of 0.0001. After training, both models were saved and used for inference.

3. Experiments

3.1 Question Answering

The Text-to-Text Transfer Transformer (T5) is a model based on the encoder-decoder architecture. It was pre-trained using teacher forcing on the diverse C4 dataset, which has achieved state-of-the-art results on benchmark NLP tasks, such as summarization, question-answering, and text classification (Raffel et al., 2020). This research evaluates the T5 large model both on question generation and question answering tasks.

Also, this research carried out experiments on developing multilingual QA systems. The traditional approaches to multilingual QA systems are highly dependent on the performance of the machine translation systems. This is because the query is first translated and then handled as a monolingual QA problem. This is susceptible to ambiguities arising from machine translation (Jiang et al., 2020). Therefore, we experimented with the MT5 model, which is a variation of the T5 model that supports multiple languages. It is pre-trained using the MC4 dataset, which contains text from 101 different languages.

BERT is a deep bi-directional Transformers-based language representation model. Unlike the T5 and mT5 model, which uses both the encoder and decoder components of the transformer model, BERT only uses the encoder component. It can be employed in different NLP tasks, such as question answering and language inference, using only one additional output layer without needing major modifications to the architecture for specific tasks.

The BERT (Devlin et al., 2019) Large uncase model is pre-trained on the BookCorpus (Zhu et al., 2015), and the English Wikipedia English Wikipedia (2023) and finetuned with the SQuAD dataset for question answering. During training BERT Large uncase model (Devlin et al., 2019), all texts are converted to lowercase, which is why the model is referred to as "uncased‘. Subsequently, these case-insensitive texts are tokenized using WordPiece with a vocabulary size of 30,000. For a single input sample, the tokenized question is added after the beginning of the sentence token, after which a separation token is added to separate the sentence from the question. The model was finetuned for 2 epochs for a learning rate of 3e-5 on the SQuAD dataset, which showed an f1 of 91.3% and an Exact Match (EM) of 86.91%.

Similarly, the T5 (Ewa S. Callahan, 2013) Large model was also finetuned on the SQuAD dataset for 5 epochs, which yielded an F1 score of 81.32 and an EM score of 77.64%. Therefore, the extractive approach using the BERT-uncased model outperformed the generative approach on the SQuAD dataset. Furthermore, during our testing, we discovered that the T5 model performed even more poorly on Wikipedia texts when compared to the SQuAD dataset. It may be due to the longer lengths of the contexts in the Wikipedia corpus compared to the SQuAD data. Moreover, when comparing the inference time of the two models on the AWS EC2 g3sxlarge instance, the BERT uncased takes 23.53% less time to answer a question compared to the T5 model.

Both the T5 model and BERT Large uncased model are not specifically trained to answer polar questions, and as a result, they did not perform well on polar questions, which require a simple "yes" or "no" answer. For this, we used a separapte model for the boolean questions. To solve this, each question was first passed to the Finetuned DistilBERT model, trained on the BookCorpus and English Wikipedia datasets to classify whether a question is polar or non-polar. This model achieved an accuracy of 99.96% on the None dataset. If the question was classified as polar, it was then passed to the Finetuned Roberta model, which enhances the BERT model using larger mini-batches and learning rates during training, to output a yes or no answer. This subsystem effectively addressed the limitations of the previous models in handling polar questions, improving the overall performance of the question generation system.

3.2 Question Generation

To generate questions, we first explored the syntactic transformation of all the sentences in the passage. Several experiments were conducted with different transformations based on our limited knowledge of English syntax. Both factoid and polar questions were generated from the syntactic transformation, as shown in Figure 4.

However, it was observed that some of the questions were not grammatically fluent, with punctuation errors and long, poorly constructed questions that lacked clear meaning and answerability. To generate more quality questions, a simple filter was implemented, but it did not significantly improve the quality of the questions generated. The filter is constructed using a matcher in spaCy Spacy API Documentation (2015). The matcher uses a rule-based approach or pattern defined by spaCy to filter sentences based on rules and patterns passed. Although the questions generated were not significantly better than the ones generated without the filter, some improvement was observed. Based on manual testing, we realized that about 30-40% of the questions generated were fluent and answerable. Furthermore, after applying the filter, there were no repeated questions generated. Subsequently, a T5 model trained on the SQuAD dataset and syntactic transformation was used to generate factoid and polar questions, respectively. However, questions generated by the syntactic transformation approach had numerous grammatical errors and were not fluent.

This research also experimented with only neural models to generate questions. The SQuAD dataset was used to train the T5 model to generate factoid questions, whereas BoolQ was used to train the T5 model to generate polar questions. Initially, the entire context was provided to the model while performing inference. However, this approach resulted in generating only a limited number of questions. Despite adjusting the maximum length parameter to tackle this issue, it did not lead to an increase in the number of questions generated during inference. To address this problem, the context was divided into paragraphs and passed separately to the model. This approach generated more questions.

The performance of the model improved with a higher number of epochs. However, due to time constraints, we trained the factoid T5 model for the or 5 epochs with a learning rate of 0.0001 and the polar T5 model for 4 epochs using a learning rate of 0.0001. Overall, the questions generated using this technique were more fluent and answerable than questions generated using the syntactic transformation method we explored earlier. By manually inspecting the generated questions, we found that the system produced over 95% of fluent and answerable questions from a pool of 20 questions generated. Some examples of the questions generated are shown in Figure 5.

```
Did Messi win the 2005 FIFA World Youth Champi
Did Messi relocate to Spain?
Did Messi struggle with injury?
Where did Messi relocate?
Did Messi first uninterrupted campaign come in
Did Messi personal best campaign to date am th
Who did Messi regain?
What did Messi won?
```

Figure 4: Snapshot of some sample questions generated by the syntactic transformation approach.

```
Did lionel messi play in the 2014 world cup?
Did messi's father have growth hormone?
Did messi ever win a cup with barcelona?
When was Lionel Messi born?
How many times has Messi won the FIFA Ballon
What is the Guinness World Records for most
When did Messi make his competitive debut?
```

Figure 5: Snapshot of some sample questions generated by the T5 neural model.

4. Discussion

4.1 Question Answering

The study compared the performance of two models, BERT-uncased and T5-Large, on the SQuAD dataset for extractive question-answering. The BERT-uncased model achieved higher F1 and EM scores than the T5-Large model, with scores of 91.3% and 86.91% compared to 81.32% and 77.64%, respectively. Additionally, the study found that the BERT-uncased model had a shorter inference time, taking 23.53% less time to answer a question compared to the T5-Large model. It is worth noting that the T5-Large model performed poorly on Wikipedia texts, which could be due to longer context lengths in the corpus.

Furthermore, the Multilingual Question Answering (MLQA) (Lewis et al., 2020) dataset was used to finetune the MT5 pre-trained model and investigate the performance and viability of this approach. In the subset of the dataset used, the question and answer are in English, while the context is in the low-resource language. The model is trained and evaluated on pairs of languages, between English and (Vietnamese, Arabic, and Spanish). Each model is trained for 5 epochs. A summary of the performance is summarized in Table 1. On each set, the model is trained on the multilingual MLQA and evaluated on the validation set.

Table 1: Result of training the MT5 model on the MLQA corpus

Translation	F1	EM	Size of Dataset
Eng-Vietnamese	22.50	6.50	6.01k
Eng-Arabic	25.14	14.51	5.85k
Eng-Spanish	39.41	26.75	5.75k
Eng-Vietnamese	22.50	6.50	6.01k

It can be observed, from Table 1, that the highest result was obtained when answering questions using Spanish contexts, while Vietnamese gave the lowest scores. The scores for these languages were low compared to the monolingual QA system. However, it shows that questions in English can be answered using passages from other languages, which is the opposite of the approach used in (Asai et al., 2021). These results show promising prospects considering the small training sets and few epochs used in training the models. However, this mode of question-answering should be further investigated to see how the model performs with more data and longer training times.

4.2 Question Generation

We could not improve the syntactic transformation because of insufficient linguistics knowledge. Furthermore, the team implemented a simple filter using matcher in spaCy and conditional filters to identify the Wh-and polar questions; however, this did not give the best results. High-precision transformation templates based on English lingual rules are required to improve the syntactic transformation system. In conjunction with the syntactic transformation approach, a T5 model, trained on the SQuAD dataset, was used to generate Wh-questions. The syntactic transformation complements the T5 model to generate polar questions. However, this approach had numerous grammatical errors and was not fluent. Therefore, a T5 model was trained on the BoolQ dataset to generate polar questions, and another T5 model was trained on the SQuAD dataset to generate factoid questions. This approach, which relied solely on T5 models, was more effective in producing accurate and fluent questions in comparison with the initial method.

Implementing this approach resulted in the desired outcome for factoid and polar questions.

5 . Conclusion

This research explored neural models for QA and syntactic transformations and neural models for QG. The QA system leverages two different models for question answering based on the type of question. To classify whether a question is polar or non-polar, a Finetuned DistilBERT model trained on BookCorpus and English Wikipedia is used. If the question is polar, the system passes the question and context to a Finetuned Roberta model, which provides a yes or no answer. For non-polar questions, the BERT large model, which is pre-trained on BookCorpus and English Wikipedia and finetuned on the SQuAD dataset for question-answering, is utilized. We achieved an f1 score of 91.3% and an Exact Match (EM) of 86.91% with the BERT large model. An experiment was also conducted on the multilingual MT5 model using English-to-Vietnamese, English-to-Arabic, and English-to-Spanish translation datasets. The study showed promising results for answering questions in English using context in other languages. Although the performance was lower than that of monolingual QA systems, further investigation is needed to see the viability of this approach.

Moreover, the research attempted to use a syntactic transformation approach for generating factoid and polar questions. Challenges arose due to insufficient linguistics knowledge, leading to unsatisfactory results. A simple filter using matcher in spaCy and conditional filters were implemented to improve the quality of the questions, but it did not yield the desired outcome. A T5 model, trained on the SQuAD dataset for generating factoid questions, was incorporated into the syntactic transformation system to enhance the performance. However, this approach had numerous grammatical errors and lacked fluency. Two separate T5 models were trained: one on the BoolQ dataset for polar questions and another on the SQuAD dataset for factoid questions to address this problem. This approach, relying solely on T5 models, proved to be more effective in generating accurate and fluent questions compared to the initial method. It was manually evaluated and found to produce over 95% fluent and answerable questions out of a pool of 20 generated questions. The findings highlight the importance of linguistic knowledge in improving syntactic transformations and demonstrate the effectiveness of T5 models trained on specialized datasets for generating different types of questions.

6. References

Kiran Adnan and Rehan Akbar. 2019. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *International Journal of Engineering Business Management*, 11:1–23.

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual Open-Retrieval Question Answering. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*:547–564.

Akshay Chaturvedi. 2018. CNN for Text-Based Multiple Choice Question Answering. *Computer Vision and Pattern Recognition Unit, Indian Statistical Institute*:272–277.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm):4171–4186

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:1342–1352.

Ankit Dwivedi and Arunothia Marappan. 2009. A Comparative Study of Neural Question Generation Models.

Susan C. Herring Ewa S. Callahan. 2013. Cultural Bias in Wikipedia Content on Famous Persons. *Journal of the American Society for Information Science and Technology*, 64(July):1852–1863

Alexander R. Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*:4508–4513

Tianyong Hao, Xinxin Li, Yulan He, Fu Lee Wang, and Yingying Qu. 2022. Recent progress in leveraging deep learning methods for question answering.

Neural Computing and Applications, 34(4):2765–2783

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual Information Retrieval with BERT.

Nguyen-Thinh Le, Nhu-Phuong Nguyen, Kazuhisa Seta, and Niels Pinkwart. 2014. Automatic question generation for supporting argumentation. *Vietnam Journal of Computer Science*, 1(2):117–127

Sweta P. Lende and M. M. Raghuwanshi. 2016. Question answering system on education acts using NLP techniques. *IEEE WCTFTR 2016 - Proceedings of 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare*:0–5

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*:7315–7330.

Bang Liu, Haojie Wei, Di Niu, Haolan Chen, and Yancheng He. 2020. Asking Questions the Human Way: Scalable Question-Answer Generation from Text Corpus. *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*:2032–2043

Nikahat Mulla and Prachi Gharpure. 2023. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, 12(1):1–32.

Yuriy Ostapov. 2011 Question Answering in a Natural Language Understanding System Based on Object – Oriented Semantics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent Advances in Neural Question Generation. (3).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67

Carol Rivas, Daria Tkacz, Laurence Antao, Emmanouil Mentzakis, Margaret Gordon, Sydney Anstee, and Richard Giordano. 2019. Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. *Health Services and Delivery Research*, 7(23):1–160..

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. :2–6.

Iqbal H. Sarker. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6):1–20.

Andrea Varga and La Ha. 2010. A question generation system for the qgstec 2010 task b. *Proceedings of the Third Workshop on Question Generation*(June 2010):80–83.

Ruqing Zhang, Jiafeng Guo, Lu Chen, Yixing Fan, and Xueqi Cheng. 2022. A Review on Question Generation from Natural Language Text. *ACM Transactions on Information Systems*, 40(1):1–43..

Ming Zhou, Nan Duan, Shujie Liu, and Heung Yeung Shum. 2020. Progress in Neural NLP: Modeling, Learning, and Reasoning. *Engineering*, 6(3):275–290.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 International Conference on Computer Vision, ICCV 2015:19–27.

English Wikipedia (2023) Wikipedia.

Wikimedia Foundation. Available at: https://en.wikipedia.org/wiki/English_Wikipedia (Accessed: May 2, 2023).
Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 International Conference on Computer Vision, ICCV 2015:19–27.
Spacy API Documentation (2015) *Matcher*.

Available at: https://spacy.io/api/matcher (Accessed: May 2, 2023).

Question Classifier V2 (2023) *alangpp255/Question_classifier_V2 · Hugging Face*. Available at: https://huggingface.co/alangpp255/Question_classifier_V2 (Accessed: May 2, 2023).