Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-15191-6

Automatic generation of short-answer questions in reading comprehension using NLP and KNN

Lala Septem Riza[1] · Yahya Firdaus[1] · Rosa Ariani Sukamto[1] · Wahyudin[1] · Khyrina Airin Fariza Abu Samah[2]

Received: 17 January 2022 / Revised: 14 December 2022 / Accepted: 30 March 2023

Abstract
In general, making evaluations requires a lot of time, especially in thinking about the questions and answers. Therefore, research on automatic question generation is carried out in the hope that it can be used as a tool to generate question and answer sentences, so as to save time in thinking about questions and answers. This research focuses on automatically generating short answer questions in the reading comprehension section using Natural Language Processing (NLP) and K-Nearest Neighborhood (KNN). The questions generated use article sources from news with reliable grammar. To maintain the quality of the questions produced, machine learning methods are also used, namely by conducting training on existing questions. The stages of this research in outline are simple sentence extraction, problem classification, generating question sentences, and finally comparing candidate questions with training data to determine eligibility. The results of the experiment carried out were for the Grammatical Correctness parameter to produce a percentage of 59.52%, for the Answer Existence parameter it yielded 95.24%, while for the Difficulty Index parameter it produced a percentage of 34.92%. So that the resulting average is 63.23%. So, this software deserves to be used as an alternative to automatically create reading comprehension questions.

Keywords Automatic Question Generation · Natural Language Processing · K-Nearest Neighbor · Text Processing · Machine Learning

1 Introduction

There are many existing English language proficiency evaluation systems, such as Test of English as a Foreign Language (TOEFL), Test of English for International Communication (TOEIC), International English Language Testing System (IELTS),

✉ Lala Septem Riza
lala.s.riza@upi.edu

[1] Department of Computer Science Education, Universitas Pendidikan Indonesia, Bandung, Indonesia

[2] College of Computing, Informatics and Media, Universiti Teknologi MARA (UiTM) Melaka Branch, Melaka, Malaysia

and others, which aim to measure a person's ability to speak English, which can later be used for academic or professional needs. In particular, IELTS is designed to assess the readiness of participants in terms of studying or practicing further education courses or higher in a university. In making the questions for the IELTS test, there are six stages, including commissioning, pre-editing, editing, pretesting, standard fixing, and test construction and grading. Of course, in making IELTS test questions, it also requires experts, costs and a lot of time. As said by Cotton [9], that in making questions it takes 50% to think about a set of questions from the total time of making. The IELTS test has four sections, namely listening, reading, writing, and speaking. In the reading comprehension test, the IELTS has several types of questions, including multiple choice, identifying information, identifying writer's views/claims, matching information, matching headings, matching features, matching sentence endings, sentence completion, summary completion, note completion, table completion, flowchart completion, label completion diagrams, and short-answer questions. Therefore, this research focuses on developing applications that are able to generate questions automatically with the type of short answer questions in the reading comprehension section. The flow of the computational model of this system was adopted from previous research, namely the model for generating What, Where, When, Who, Which, and How (5W + 1H) questions [2].

The type of short answer question on IELTS is the type that asks the test taker to answer questions related to factual information about the details contained in the text. Test takers are required to write down their answers in the form of words and numbers on the answer sheet provided, and write their answers using words from the reading text. In addition, there are usually instructions regarding the length of the answer that can be written, for example "No More Than Three Words and/or A Number from the passage", "One Word Only", or "No More Than Two Words". If the participant writes the answer beyond what is instructed, then the participant will lose points for that question. Numbers can be written in the form of words or numbers, and words written using hyphens will be considered as single words. The questions are in the same order as the information in the text [25]. This type of question aims to assess the ability of test takers to find and understand the right information in the text.

Systems that are able to generate questions automatically have been widely used in previous studies. For example, research conducted by Mazidi and Tarau [29], which discusses automatic question generation using the DeconStructure algorithm, dependency, SRL parse, TextRank algorithm, and internal NLU analysis methods. Meanwhile, the research conducted by Kumar et al. [22] used the Part-of-Speech (POS) Tagger and Support Vector Machine (SVM) methods to generate fill-in-the-blank questions. The tools he uses are of course different, including "Amazon Mechanical Turk", "scikit-learn" python package, "Radial Basis Function (RBF) kernel", "WordVec", and "WordNet". The result of his research is a system called "RevUP". Furthermore, research related to automatic question generators uses semantic pattern recognition to create questions with different depths and types [28]. This research uses the method of Negation Detection, and Linguistic Considerations, with tools "SENNA" software, Python, and "WordNet". The results of his research show a 44% reduction in error rates relative to the previous best system, top average across all metrics, as well as a 61% reduction in error rates on grammatical assessments.

This study seeks to produce an automatic question generating system with the type of questions generated in the form of short answer questions in reading comprehension using NLP and the KNN. The NLP method is used to process data in the form of text

while KNN, which is a machine learning method, is used to choose the best question based on training data (i.e., data on questions that have been raised in IELTS questions). The machine learning method is intended so that the resulting questions have a quality that is not much different from the question data that has appeared before. The main contributions of this research are as follows: (i) by performing machine learning method we attempt to maintain question quality from datasets of historical questions; (ii) questions can be generated automatically and fast because the input data are from articles and users just need to determine numbers of questions; and (iii) this research works on Part of Speech (POS) tagging instead of using words, so the model constructed can be used for all words available in dictionary.

2 Research method

2.1 Computational model

Figure 1 shows the flow model of the system built in this study; this path was adopted from research conducted by Ali [2]. The system built in this study can do scrapping or retrieve article content from the website, by entering the URL of the website address. In doing this scrapping using the library provided by python, namely "newspaper3k". This library can be used to retrieve the content, author, and publish date of articles. Before the data of this article is extracted to become a candidate question, previously this data must be processed so that it can be used to extract questions. A detailed explanation of this computational model is as follows:

2.2 Data collection

This stage is needed to collect articles that will be extracted into several questions that are suitable to be used as short answer questions in IELTS reading comprehension. Fig. 2, shows an example of an article taken from the BBC news website, which is about the corona situation in India. The library used at this stage is "newspaper3k", which is to retrieve article contents from links entered by users.

2.3 Data preprocessing

After obtaining the article that will be converted into several questions, the next step is to separate the sentences. This separation is done with the condition that the beginning of the sentence must begin with a capital letter and end with a period, if it does not meet the requirements then the sentence will not be processed to the next step. Figure 3, shows an example of the result of separating sentences from paragraphs.

The next step after splitting sentences is to preprocess each sentence generated from the paragraph. This preprocessing is carried out with the aim of eliminating characters and symbols, so that they can be processed at the next stage. Removing these characters and symbols uses the "regex" function, and an example of the results of the preprocessing can be seen in Fig. 4.

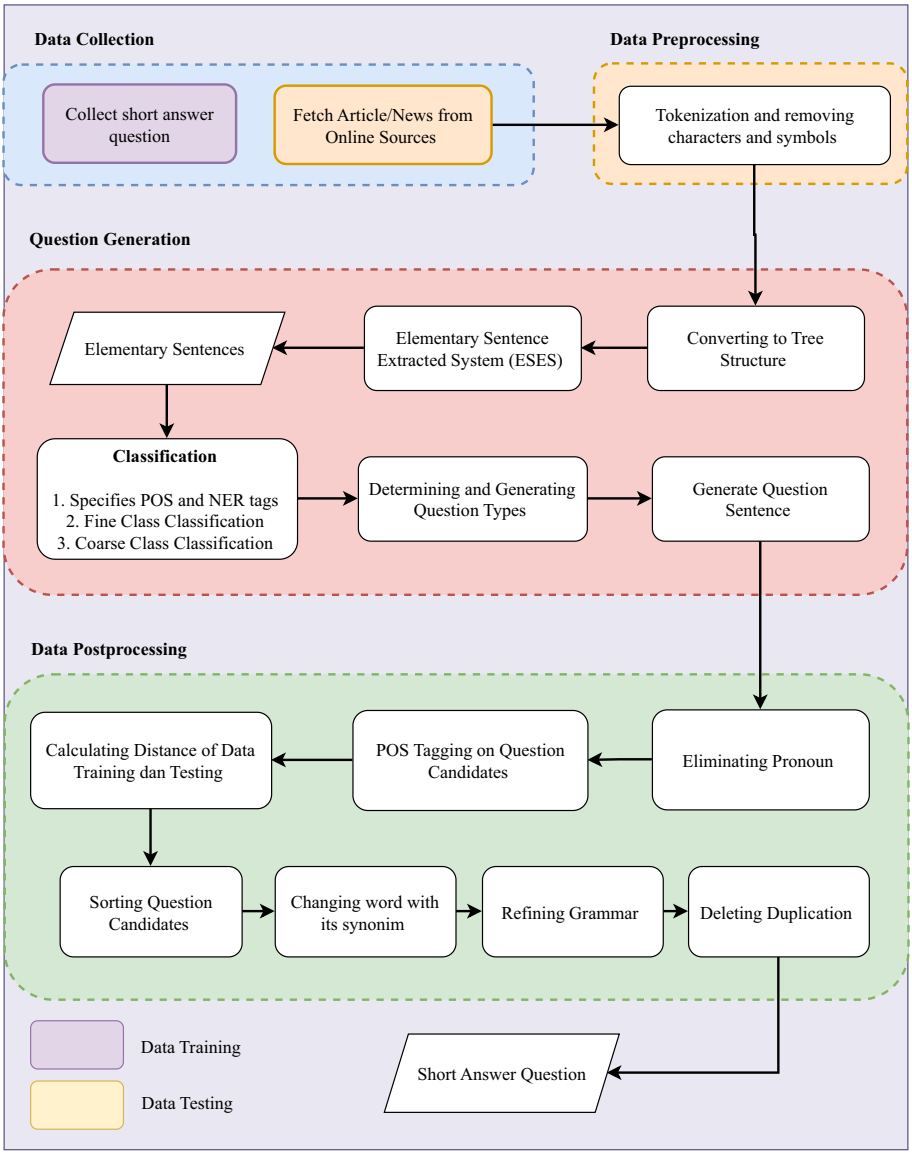


Fig. 1 Automatic question generation model flow

2.4 Converting to tree structure

At this stage, sentences that have been cleaned of characters or symbols using the regex of the previous stage will be converted into a tree structure, because characters or symbols cannot be converted to POS tags, so "regex" is needed. The library used to perform this conversion is "Stanford Core NLP" via "NLTK". The results of the conversion into a tree will be used by the next stage, namely to extract simple sentences from complex

More than two million Indians have now tested positive for Covid-19, acconting to official figures. The country confirmed the last million cases in just 20 days, faster than the US or Brazil which have higher numbers. Testing has been expanded considerably in India in recent weeks but the situation varies across states. Spurred by a low death rate, the nation continues to reopen even as new hotspots drive the surge in cases. But some states have imposed restrictions. The recent measures include local, intermittent lockdowns, sometimes limiting activity in specific cities or districts. India is now the third country to cross the two million mark. It reported 62,170 cases in the past 24 hours, taking its total tally up to 2,025,409. The country has reported around 40,700 deaths so far. While that is the world's fifth-biggest total, experts say it is not very high given the country's population of 1.3 billion.

Fig. 2 Examples of news articles taken from the website

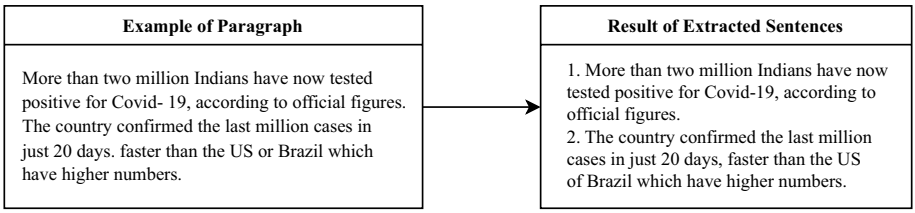


Fig. 3 Breaking paragraphs into sentences

sentences. An example of the results of the conversion into a tree structure can be seen in Fig. 5. Previously the results of this conversion were in the form of a data tree, therefore the conversion process from tree data types to strings was required.

2.5 Elementary sentence extracted system (ESES)

Elementary Sentence Extracted System (ESES) is a system used to extract elementary sentences or simple sentences from complex sentences. Because in making questions, simple sentences are needed, ESES will be very useful in terms of extracting these sentences. Figure 6 shows the flow of the computational model for extraction of simple

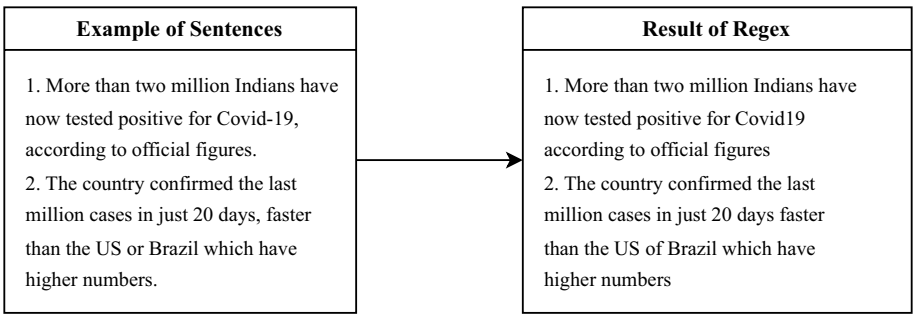


Fig. 4 Remove characters and symbols

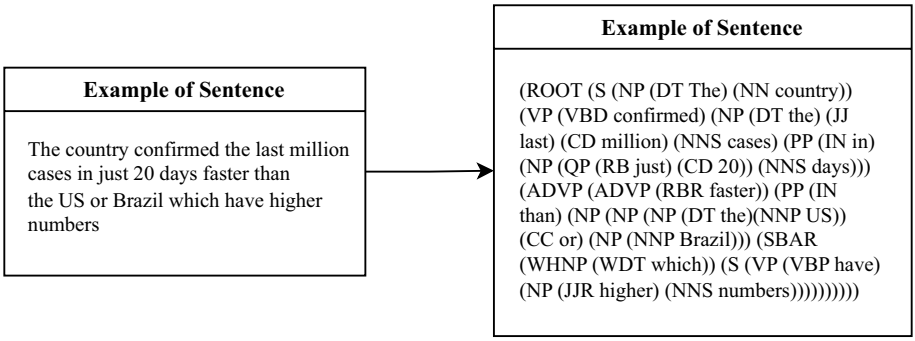


Fig. 5 Converting to tree structure

sentences which is divided into four arrays, including NP or noun phrase, VP or verb phrase, word depth in the sentence, and also the word order in the sentence.

Then each NP and VP will be combined by looking at the depth and order of each NP and VP. A simple sentence has only one subject, no less and no more, meanwhile, a sentence that is considered a complex sentence is a sentence that has more than one subject. This requirement is in accordance with the research conducted by Kalady et al. [20]. Figure 7 shows an example of the result of extracting simple sentences from complex sentences.

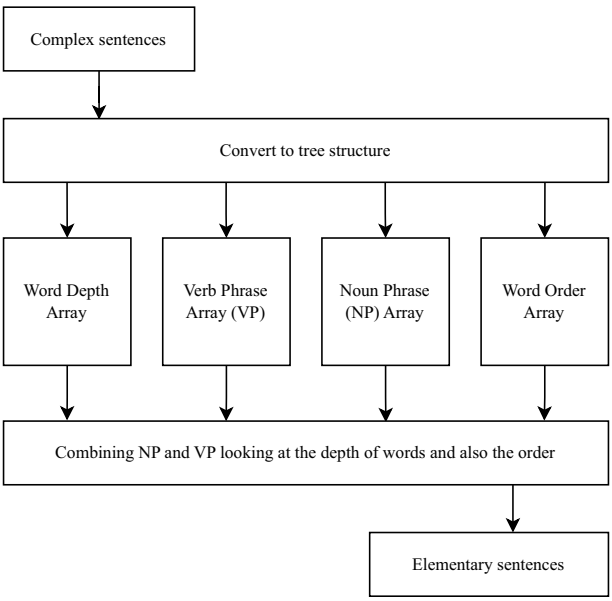


Fig. 6 ESES flow model

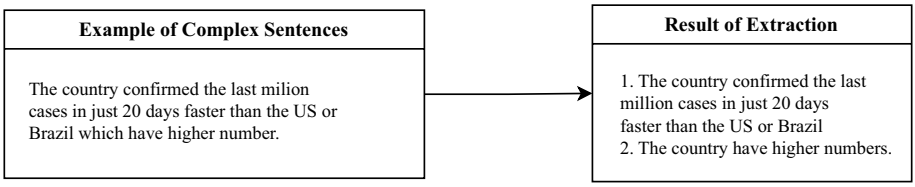
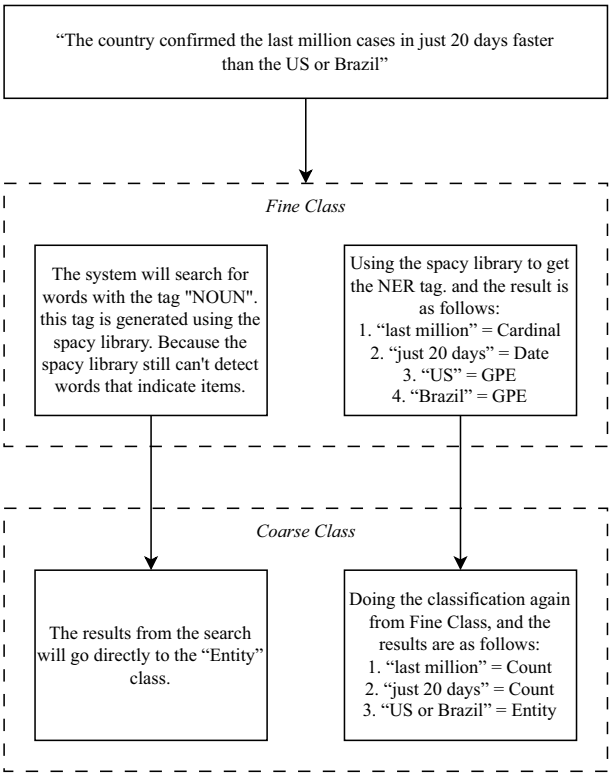


Fig. 7 Extraction of complex sentences into simple sentences

2.6 Classification

After getting simple sentences from the previous stage, the next stage is to determine the Fine Class and Course Class. As seen in Fig. 8, Fine Class determines the initial class of each word in the sentence, while Course Class is a regrouping of Fine Class which is divided into five predetermined classes, namely Human, Location, Entity, Time, and Count. These five classes are the rules of the sentence, such as "Human verb Human", according to the coarse class produced from the sentence.

Fig. 8 Classify to fine and coarse class



2.7 Determining and generating question types

The sentence rules resulting from the previous stage will be used at this stage to determine the types of questions that can be generated. Then, this stage will also produce a question sentence and the appropriate answer from the sentence. For example, in Fig. 9, the sentence rules will be extracted from simple sentences, which will then obtain the types of questions by looking at the subject, object, and preposition. Because the sentence rules have "time" then one of the types of questions that can be generated is "when". What types of questions can be generated with sentence rules.

After getting what types of questions can be generated in one sentence, the next step is to make question sentences according to the type of question. Each type of question will produce one question, except for questions with the type of "whom" where in the sentence there is a coarse class with more than one type of "Human". For example, it can be seen in Fig. 10, which is from the type of question "when" then this will produce a question sentence in the form of "When Husam cooks the rice?", this question asks for the time or according to the rules of the sentence "Time", with the answer "at 4 pm".

2.8 Eliminating pronoun

The resulting question sentences still have to be processed by sorting or cleaning the question sentences and answers from pronouns, because with the question sentences and answers that have pronouns it will cause ambiguity in answering the question. Therefore, at this stage we will remove candidate questions that have pronouns in the question sentence or in the answer. For example, it can be seen in Fig. 11, in the sentence the question has a pronoun, namely "The country", therefore the question will be deleted and will not be processed at a later stage.

2.9 POS Tagging on question candidates

The next step is to convert it into a POS tag and determine the number of words in the question candidate sentence. POS tag data and many words in this question sentence are useful for determining the feasibility of questions from candidate questions generated by the system.

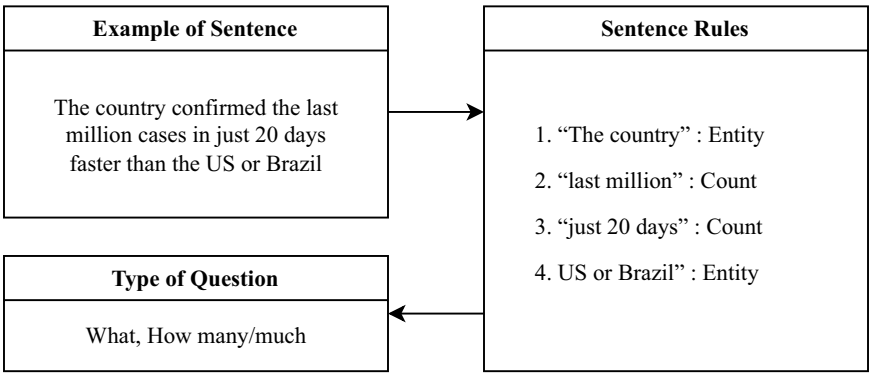


Fig. 9 Getting the question types

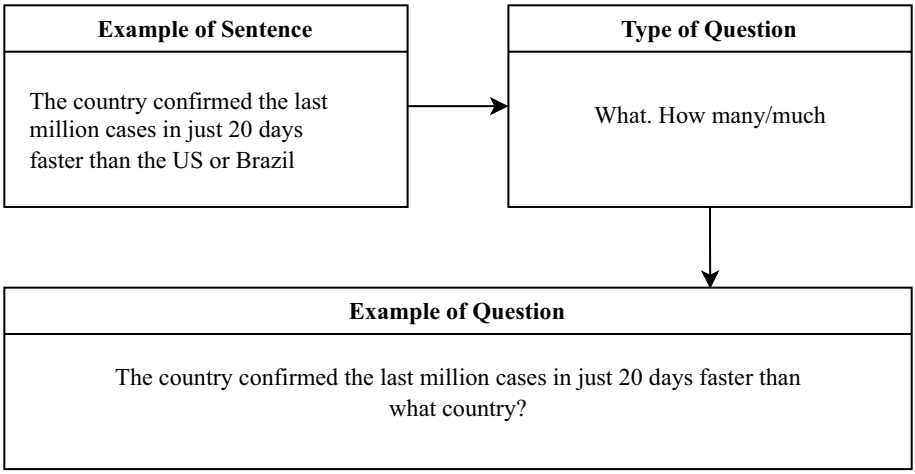


Fig. 10 Generate a question from a question type

The steps of the POS tag and the determination of the number of words in a sentence can be seen in Figs. 12 and 13. From the sample sentence questions, preprocessing was carried out to remove characters and symbols, after which they were converted to get a POS tag, which finally obtained many words in a sentence by counting the number of words. POS tags. The library used at this stage is "Stanford Core NLP", which is to convert sentences into trees.

2.10 Calculating distance of data training dan testing

The next step is to convert it into a POS tag and determine the number of words in the question candidate sentence. POS tag data and many words in this question sentence are useful for determining the feasibility of questions from candidate questions generated by the system. From the sample sentence questions, preprocessing is carried out to remove characters and symbols, after that they are converted to get POS tags, which finally get a lot of words in the sentence by counting the number of POS tags. Before using the KNN formula, the POS tag is first converted to a numeric value. The first step is to initialize each tag into a number, the numbers for each tag can be seen in Table 1.

After getting the value of each tag, the next step is to determine the value of S, provided that the range is from 0 to 100 with 36 tags. The calculation can be seen in Eq. 1, so that the S value is 2.86. Since the value of S has been obtained, the next step is to calculate the

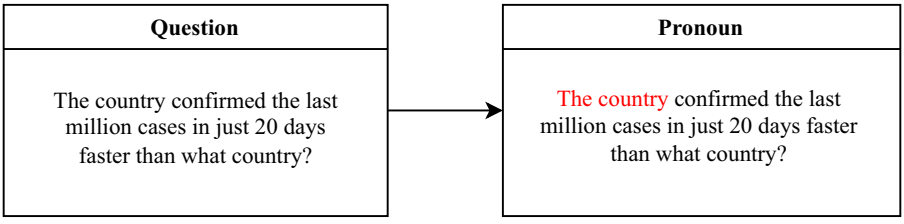


Fig. 11 Examples of question sentences that have a pronoun

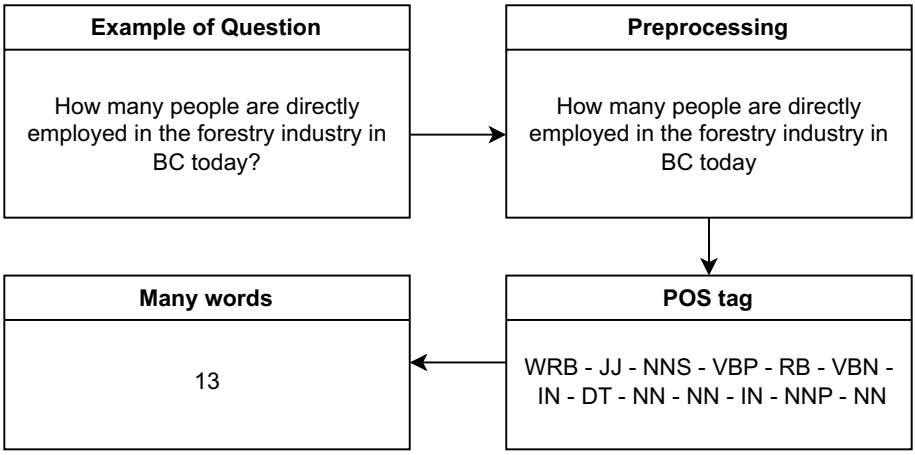


Fig. 12 The results of pre-processing training data, POS tag conversion, and many words

value of V, which is the numeric value of each tag. As seen in Eq. 2, the value of V is the value of S multiplied by the tag value and then subtracted by one.

$$S = \frac{100}{36 - 1} = 2.86 \tag{1}$$

$$\begin{aligned} &0 = -1, P = tag\ value \\ &V = (2.86 * P + O), tag = VB = 27 \\ &V = (2.86 * 27 + (-1)) = 76.22 \end{aligned} \tag{2}$$

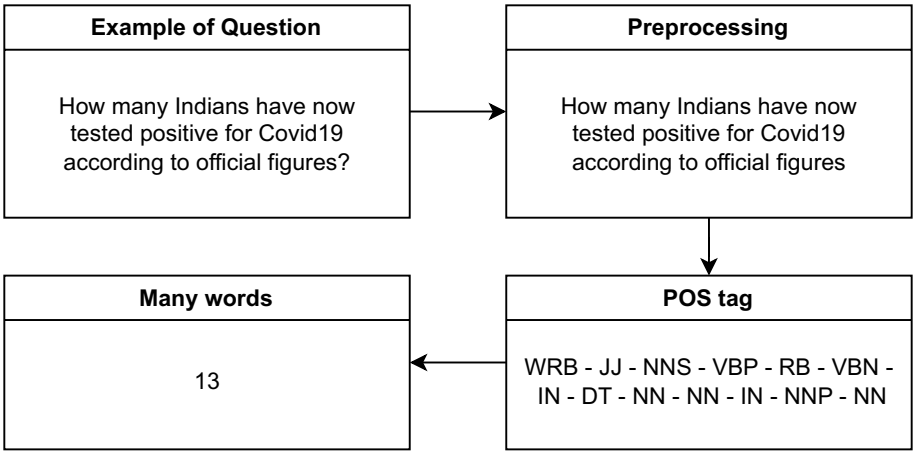


Fig. 13 Preprocessed candidate query results, POS tag conversions, and word count

Table 1 Initialize tag values

Tag	Value	Tag	Value	Tag	Value
CC	1	NNS	13	TO	25
CD	2	NNP	14	UH	26
DT	3	NNPS	15	VB	27
EX	4	PDT	16	VBD	28
FW	5	POS	17	VBG	29
IN	6	PRP	18	VBN	30
JJ	7	PRP$	19	VBP	31
JJR	8	RB	20	VBZ	32
JJS	9	RBR	21	WDT	33
LS	10	RBS	22	WP	34
MD	11	RP	23	WP$	35
NN	12	SYM	24	WRB	36

After the numerical value of the POS tag is obtained, it is possible to calculate the distance between the test data and the training data. Figure 14 is an example of calculating the distance between training data and test data, the result of this calculation is 91.96, where the smaller the number, the more similar the test data to the training data. Because the results are 91.96, it can be said that the test data questions are not similar to the training data questions.

2.11 Sorting question candidates

After getting a list of question sentences that have been sorted from question sentences and also answers that have pronouns and getting the value of the distance between the training data and the test data, the next step is to sort the question sentences according to the distance from the training data. In order to obtain a list of candidate questions that are sequentially according to their proximity.

2.12 Changing word with its synonim

Next is the stage of changing some words that have been determined with their synonyms. The words in the sentence that will be replaced with synonyms include words that have adjective, adverb, and verb tags. The conversion of certain words into their synonyms is intended to make the resulting question sentences more difficult than the original sentences whose words were taken from the article text. Therefore, this stage is necessary so that the resulting candidate questions have a more difficult level of difficulty than before. For example, Fig. 15 shows that the word "now" is changed to its synonym "today", and the word "tested" becomes "proved".

2.13 Refining grammar

This stage is to check the grammar of the question candidate sentence, which is to make corrections to the grammatical errors in the question candidate sentence. To check this grammar, we use a library in the python programming language, namely the language

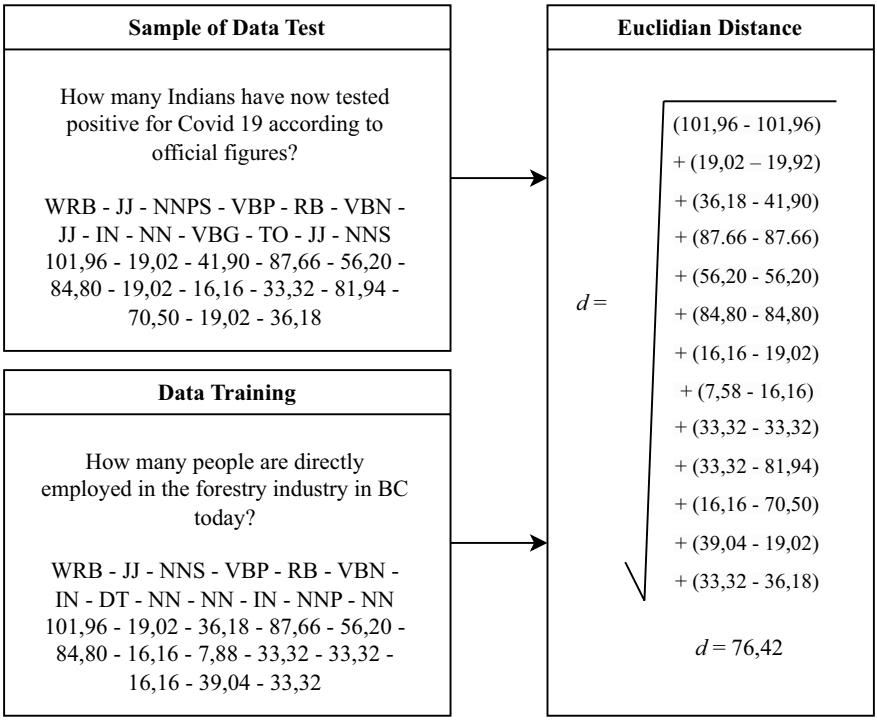


Fig. 14 Example of the distance between the training data and the test data

tool. For example, Fig. 16 is the result of checking using the language tool library. The change is that what was originally "ebooks" was changed to "e-books".

2.14 Deleting duplication

This stage is the selection or discarding of candidate questions that come from the same sentence. An example of the final result of this generated question can be seen in Fig. 17.

Evaluation Methods.

The following is methods to be used to evaluate the results:

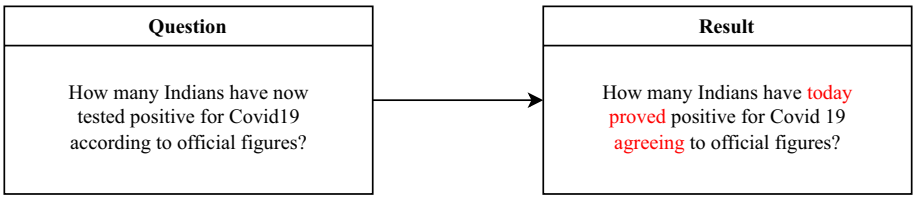


Fig. 15 An example of converting certain words to their synonyms

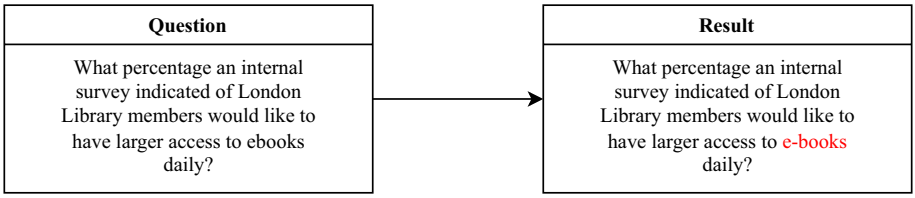


Fig. 16 Example of grammar checking and improvement

Grammar Checker: This analysis is used to find out grammatical errors in question sentences. We use the help of a website application to check grammar, by using a website with the url https://www.reverso.net/spell-checker/english-spelling-grammar/.

Expert Judgement: The next stage is to evaluate the feasibility of the questions by human experts. It is intended that the resulting questions can be assessed for feasibility by considering the following criteria:

Grammatical Correctness (GC): that is to determine whether the question sentences generated by the system are syntactically well-formed. At this point, we divide 3 levels of value according to the number of grammar errors, including:
Best: Generated questions do not have grammatical errors.

More than two million Indians have now tested positive for Covid- 19, according to official figures. The country confirmed the last million cases in just 20 days, faster than the US or Brazil which have higher numbers. Testing has been expanded considerably in India in recent weeks but the situation varies across states. Spurred by a low death rate, the nation continues to reopen even as new hotspots drive the surge in cases. But some states have imposed restrictions. The recent measures include local, intermittent lockdowns, sometimes limiting activity in specific cities or districts. India is now the third country to cross the two million mark. It reported 62,170 cases in the past 24 hours, taking its total tally up to 2,025,409. The country has reported around 40,700 deaths so far. While that is the world's fifth-biggest total, experts say it is not very high given the country's population of 1.3 billion

Questions 1-2
Answer the questions below.
*Choose **NO MORE THAN THREE WORDS AND/OR A NUMBER** from the text for each answer*

Write your answers in boxes 1-2 on your answer sheet
1. What country is today the 3rd country after crossing the two million mark?
2. How many Indians have today proved positive for Covidl9 agreeing to offical figures?

Answers
1. India

2. More than two million / More than 2.000.000

Fig. 17 Example of paragraphs and system generated questions

Good: Generated questions have one or two grammatical errors.
Worst. Generated questions have more than 3 grammatical errors.
Answer Existence (AE): namely determining whether the answer to the question is appropriate, or the question can be answered by the answer key correctly.
Difficulty Index (DI): which is to determine the value of how difficult the questions are generated by the system. This point the author divides into 3 levels of value, including: easy, medium, and hard.

Next, we calculate the percentage value of the expert assessment using the formula shown in Eq. 3.

$$Percentage = \frac{Total\ score\ of\ the\ evaluation}{Maximum\ score} * 100\% \quad (3)$$

3 Experimental design

In this study, the data used is news article data that is used to generate questions from the article, which will then become test data or candidate questions. Other data is IELTS question data that has appeared which is used for training data. The number of training data sources used are 49 books and 3 websites with a total of 1010 questions. Examples of e-books taken as data collection for IELTS training questions include:

Cambridge Practice Tests for IELTS 2 [6]
Cambridge Practice Tests for IELTS 1 [31]
Cambridge IELTS 3 With Answer Edition [7]
IELTS Reading Tests [30]
Collins Reading for IELTS [41]
Insight into IELTS Update Edition [18]
Barron's IELTS [26]
Headway Academic Skills: Level 1 Student Book [15]
Master IELTS 6: IELTS Precise Reading [35]
High Impact IELTS Academic Module [5]

As for the data used as candidate questions, it was obtained from several trusted websites, namely the BBC, CNN, The Jakarta Post, and the New York Times. The topics of each article used are different, including health, hoaxes, holidays, the environment, and sports (See Table 2).

4 Results and discussion

The results of this experiment are questions and answers generated by the developed system. Table 3 shows the candidate questions and answers generated by the system in this experiment. As an example, some questions generated by the system are taken. Grammar analysis is used to find out grammatical errors in question sentences. It aims to find out whether the question sentences generated by the system are in accordance with the

Table 2 Article sources

No. Text	Link	Website Name	Topic
1	https://www.bbc.com/news/world-asia-india-53674857	BBC	Health
2	https://edition.cnn.com/2020/08/06/world/fake-beirut-video-missile-intl-trnd/index.html	CNN	Hoax
3	https://www.thejakartapost.com/life/2020/10/02/ebook-loans-book-dispensers-how-are-libraries-adjusting-to-the-pandemic.html	The Jakarta Post	Book
4	https://www.thejakartapost.com/life/2020/10/02/from-athens-to-kyoto-try-coffee-like-a-local.html	The Jakarta Post	Food
5	https://www.nytimes.com/2020/10/16/technology/twitter-new-york-post.html	New York Times	Technology

Table 3 Generate question experiment results

No.	Questions	Answers	Text
1	What country is today the 3rd country after crossing the two million mark?	India	1
2	What country is also getting the highest number of daily newly cases in the world?	India	1
3	How many Indians have today proved positive for Covid19 agreeing to official figures?	More than two million	1
4	Case numbers are emerging quickly for instance in the southerly state of what?	Andhra Pradesh	1
5	The country has covered how many deaths then far?	40,700	1
6	What are already spreading on every Major social media platform?	Doctored videos	2
7	Who is a producer is the Facebook page of Beirut based CNN Arabic social media?	Mehsen Mekhtfe	2
8	In what city fixed videos of Tuesdays devastating explosion are already spreading on every Major social media platform?	Beirut	2
9	What city where the 72 municipal libraries have stayed close for months?	Los Angeles	3
10	What are even limited due to the healthful guidelines in effect in the UK?	London Library	3
11	What percentage an internal survey indicated of London Library members would like to have larger access to e-books daily?	60 percent	3
12	What situation has reinforced the importance of online content at a time when accessing forcible collections?	Coronavirus	3
13	What should so be stewed double in a row?	The coffee	4
14	What is always serviced three times with water brought each time?	The coffee	4
15	What had from brass in The typical Turkish coffee pot foreboded or copper?	the cezve	4
16	Water is so brought to seethe up the potion over what?	a charcoal fire	4
17	The typical what coffee pot is a crucial piece of kit ?	Turkish	4
18	Trump utilizes social media as what?	a megaphone	5
19	The companies prohibited candidates from taking what?	an election victory	5
20	From the start what article was problematic?	the New York Post	5
21	Some people are already utilizing what for call for election violence?	the sites	5
22	What country operatives were discovered to have utilized the sites to seed discord in the 2016 election?	Russian	5

Table 4 Results of evaluation by expert

Parameter	Maximum Score	First Expert		Second Expert	
		Score	Percentage	Score	Percentage
Grammatical Correctness	63	29	46.03%	46	73.02%
Answer Existence	42	39	92.86%	41	97.62%
Difficulty Index	63	21	33.33%	23	36.51%

grammar rules or not. This error grammar checks based on error, misspelling, uncertainty, and undefined. Of these 21 questions, the grammar check resulted in a total of 5 questions that had grammatical errors, so the result was a percentage of correct grammar of 76.19%.

Next, we evaluate the feasibility of the questions by human experts. It is intended that the resulting questions have a high level of feasibility. This evaluation was carried out by two experts. Table shows the results of the evaluation carried out by the expert; it can be concluded that the first expert for the grammatical correctness parameter gave a score of 29 with a percentage of 46.03%. The answer existence parameter is given a score of 39 with a percentage calculation of 92.86%. As for the difficulty index parameter, a score of 21 is given with a percentage calculation of 33.33%. While the assessment of the second expert with a total score per parameter, namely the GC parameter has a total of 46, the percentage is 73.02%. The AE parameter has a total score of 41 with a percentage of 97.62%. The DI parameter has a total score of 23 with a percentage value of 36.51% (See Table 4).

After calculating the percentage by looking at each text and each expert, the next step is to perform calculations by combining the results of the evaluation carried out by the first expert with the second expert. This is done to be able to see the final percentage value and also its average. The results of the calculations that have been carried out can be seen in Table 5, the ideal value, the maximum score of each parameter is different, namely the grammatical correctness parameter has an ideal value of 3 with a maximum score of 63 and has a total score of 75. The answer existence parameter has an ideal value of 2 with a score of a maximum of 42 and has a total score of 80. The difficulty index parameter has an ideal value of 3 with a maximum score of 63 and has a total score of 44.

Furthermore, this study also tries to compare it with existing research, as shown in Table 6. Like the research conducted by Heilman and Eskenazi [16], the questions generated by the system have a precision of 80%, this study uses the thesaurus extraction technique method, the evaluation used is the thesaurus extraction technique. It is also used using experts in assessing the questions generated by the system. This research produces multiple choice questions. In contrast to the research conducted by Heilman and Eskenazi,

Table 5 Calculation results for each parameter

Parameter	Ideal Value	Maximum Score	Total Score	Percentage
Grammatical Correctness	3	63	75	59.52%
Answer Existence	2	42	80	95.24%
Difficulty Index	3	63	44	34.92%
Average				63.23%

Table 6 Comparisons with previous research

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
1	[2]	In this study, an automatic question generator system from sentences was developed. That will generate possible questions to use. With modules or processes in data processing	Syntactic Parsing, POS Tagging, Named Entity Tagging	WTML (Web Testing Markup Language), HTML, Question Generation Shared Task Evaluation Challenge 2010 (QGSTEC), Elementary Sentence Extraction System (ESES)	Evaluation system developed using QGSTEC. The type of question generated is 5W + 1H. The result is an average recall below 0.300	Short Answer Question
2	[29]	The author proposes a new approach to an automated question generating system that can improve the presentation of accepted questions compared to state-of-the-art systems	NLP, DeconStructure algorithm, SRL	MTurk, Python, Microsoft Research's SPLAT, H&S system	From 200 questions by comparing the system with the H&S system, the resulting average rating exceeds the Heilman & Smith system, with a value of 3.7 and while the Heilman & Smith system has an average rating of 2.9. Or 72% of the questions from the system generated results can be accepted	Multiple Choice
3	[22]	Continues to create a variety of questions from previous research, which is useful as an effective method to help students learn better	Part-of-Speech (POS) Tagger, Human Intelligence Tasks (HITs), Support Vector Machine (SVM)	Amazon Mechanical Turk, Word2Vec, WordNet, TF-IDF search engine, phyton	Generates an automatic question generation system called RevUP. With the first step is sentence selection, then gap selection, and the last is distractor selection. By using a discriminative classifier in conducting training, it produces an accuracy of 81%. With 94% distractor generated it is good	Gap Fill

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
4	[28]	The results of the learning achievement carried out by students can be seen from the answer to the question, to measure the depth of learning that students do. Therefore, researchers developed an automatic question maker system to measure student learning	semantic pattern recognition	SENNA, NLTK, phyton, H&S, LPN&W	The validation results showed a 44% reduction in error rates than the previous system, all metrics were above average, and a 61% reduction in error rates on grammatical assessments	Short Answer Question
5	[16]	The aim of this study was to create an automatic question generator with a specific type, namely a vocabulary assessment. Create a system that can become an English learning tutor	thesaurus extraction technique	WordNet, Gigaword corpus	This study produced an automatic question generating system by comparing two techniques, namely the thesaurus extraction technique and using humans. The result is that using the system get 80% precision of the questions generated	Multiple Choice
6	[42]	The researcher proposes a neural model to generate natural-language questions from documents. And explain how to train for models using a combination of supervised and reinforcement learning	Supervised, reinforcement learning	dataset SQuAD, neural machine translation (NMT), Seq2Seq System	By comparing the system that the researcher developed with the Seq2Seq system, the result is that seen from the Fluency (PPL) value, the research system has a value of 175.7 and Seq2Seq has a value of 153.2. The system developed by the researcher generates more specific questions than Seq2Seq, but there are deficiencies in the context	Short Answer Question

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
7	[37]	Creating an automatic question generation system that makes it easier to make questions with error identification type in the TOEFL	Natural Language Processing, and Levensthein Distance	StanfordCoreNLP, python, php, framework codeigniter	The results of the study showed that the quality of the questions generated was considered good, because two expert judgments gave a percentage value of 82%	Error Identification
8	[40]	Creating an automatic question generation system that makes it easier to create questions with sentence completion types in the TOEFL	Natural Language Processing, and using K-Nearest Neighbor algorithm	StanfordCoreNLP, python, php, framework codeigniter, Ultra Lingua API	The results of the evaluation carried out resulted in a percentage of 81.93%, which could be seen that the quality of the questions produced was good, with 81.25% consistency of answers and 70% of the similarity in blank positions	Sentence Completion
9	[27]	Multiple choice questions are one type of question used to measure a person's ability to understand learning. However, making multiple choices manually is a waste of time. Therefore, researchers developed an automatic question generator system	Support Vector Machine (SVM) classifier, Parse Tree Matching (PTM) Algorithm, Conditional Random Field (CRF)	WordNet, Stanford CoreNLP Suit, NER system	Produce a multiplechoice question generator system with an accuracy of 93.684% from 5 evaluators. Of the 95 sentences, evaluator 1 gives 90, evaluator 2 gives 89, evaluator 3 gives 87, evaluator 4 gives 90, and evaluator 5 gives 89	Multiple Choice
10	[3]	Assessment of students' abilities in the learning process is better using an evaluation or test system. On the other hand, evaluation design and planning requires a lot of time. Therefore, this research was conducted	Structure Mapping Theory (SMT), Vector Space Model (VSM)	Gene Ontology, People & Pets Ontology, Pizza Ontology	They proposed solving 8% of the questions generated from Gene Ontology, 67% of the questions generated from People and Pets Ontology, and 88% of the questions generated from Pizza Ontology	Multiple Choice

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
11	[17]	Creating an automatic question generator system that can be useful for conducting tests and evaluating students' language skills. With multiple choice questions and having different question levels and categories	exponential moving average (EMA), Chi-Square test	Tgrep2 patterns. Sumber teks: Time For Kids, Student Times, Voice of America, CNN, China Post Online and Yahoo! News	Generates an automatic question generator system with multiple choice question types and has different question levels and categories. The results showed that the rectification rate in the experimental group on average was significantly higher than the control group. Experimental group (M = 0.54, SD = 0.29)	Multiple Choice Gap Fill
12	[13]	Helping students in learning by creating an automatic question generator system from various texts, for example from articles, papers, and websites that are freely chosen by students. This research focuses on generating questions automatically from a sentence	NLP, k-nearest neighbor (kNN) classifier	WordNet, Penn Treebank II, Conditional Random Field (CRF), Google AJAX Search API, PHP, AJAX	By being tested by an expert to answer questions generated from the system, the text given is two texts, namely text A, text B, and text C. incorrect language. Whereas for C text, more than 60% of the tricks are not good or inappropriate by more than one expert	Multiple Choice Gap Fill
13	[24]	This study intends to create a system that can support students in learning. The purpose of this automatic question generator is a reading comprehension or vocabulary assessment	NLP, Latent Semantic Analysis (LSA)	Writers Workshop, SaK, Sourcer's Apprentice Intelligent Feedback system (SAIF), intelligent tutoring system (ITS)	Generates an automatic question generating system called G-Ask. Researchers evaluated using training data from 45 papers and 33 literature review papers to test 534 citations and 469 extraction results. Citation Extraction Rate of 88%	Short Answer Question

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
14	[8]	Create an automatic question generator from text that has topics. By using Latent Dirichlet Allocation (LDA) to identify sub-topics	Latent Dirichlet Allocation (LDA), Natural Language Processing (NLP), Natural Language Generation (NLG), Named Entity (NE), Semantic Role Labeling (SRL)	extended string subsequence kernel (ESSK), ASSERT, Word Sequence Kernel (WSK), String Subsequence Kernel (SSK)	There is baseline 1 and baseline 2 which are meant to make QG a system in two ways. The result is that at baseline 1 the score for the relevance of the topic is 2.15 and the syntax is 2.63. Whereas for baseline 2, the relevance score is 3.24 and the syntax is 3.30. And the QG system created by the researcher produced a relevance score of 3.48 and a syntax of 3.55	Short Answer Question
15	[21]	Due to the difficulty in making a question-answering (QA) system, a natural language is generated. So this study aims to provide a new style in QA architecture, by using sentences in the document as the source of QA	NLP, Named Entity Recognition (NER)	Phyton, MySQL DBMS, NER tool	The result of this research is a QA system that can make questions from paragraphs. The research methods are sentence split, NER, question generation, question filtering, and question / answer indexing	Short Answer Question
16	[36]	It is used to generate sentence completion type in TOEFL	NLP and KNN	Phyton	The validations were conducted by utilizing human expert and grammar tools	Sentence completion type
17	[19]	Reduce barriers such as financial cost or extremely long development timelines to creating courseware by automating basic steps that require significant manual labor	Supervised and unsupervised machine learning models. And a variety of syntactic and semantic NLP methods	spaCy library	SmartStart app for generate questions. SmarStart could transform a textbook into courseware	Fill-in-the-blank (FITB) and matching

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
18	[4]	Generating the factual questions from unstructured text in the English language	Combines linguistic approach based on various types of sentence patterns with machine learning approaches (multi-label classification)	POS taggers, named entity recognizers, semantic role labellers etc	The generated questions from the result outperforms the state-of-the-art systems and the questions are also comparable to questions created by humans	Short answers questions
19	[39]	Integration of and automatic question generation system and computerized adaptive test	NLP	Exford3000 words, GSL word lists	All computerized adaptive test simulations performed better than the baseline, a linear test, in estimating the test taker's true profieciency	Reading comprehension questions
20	[11]	Generating subjective questions and also and evaluation system is suggested for assessing the answers	NLP	SQuAD, RACE dataset	It uses a set of model answers taken from different textbooks and subject experts to evaluate the answers. The automated appraisal process can reduce ethe manual effort of the human	Short answers questions
21	[23]	Forming an ontology generation model and a template model for generating questions that are not domain-specific	NLP	Protégé, PHP	The results of ontology generation concerning "animal" produced 3811 axioms, 59 classes, 1118 instances, and 26 object properties. All components of the ontology were then used as the basis for generating questions	Short answers questions
22	[44]	Novel approach to automatic questions generation using semantic role labeling for morphologically rich languages is presented	Semantic role labeling, conditional random fields	Prague Dependency Treebank, WordNet, CROWN	Expert evaluation of the system showed that 68% of the generated questions could be used for educational purposes	5W + 1H questions

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
23	[34]	Presenting a pipeline for generating and evaluating questions from text-based learning materials in an introductory data science course	NLP	Python, beautiful soup library, Google's T5, SQuAD	Generated questions were rate favorably by: information score, automated rating by a trained model and manual review by human instructos	Short answers questions
24	[12]	To generate reading comprehension questions and multiple-choice questions on grammar form a given English text	Machine learning, NLP	Text-to-text transfer transformer, pre-trained natural language understanding model, SQuAD 2.0 dataset	The system takes and average time of about 1 s to generate a Wh-question and it generates a MC question almost instantly	5w + 1 h, multiple questions
25	[38]	Proposes a questions-distractor join generation framework (QDG)	NLP	RACE dataset	The model achieves a giant breakthrough in the question-distractor pair generation task	Reading comprehension
questions						
26	[43]	To assist teachers in question generating, meanwhile, to provide more CET exercises for students	NLP, attention mechanism	Seq2seq	Applied the technology of abstractive text summarization technology to the automatic generation of the topic belonging to the English reading comprehension	Reading comprehension questions
27	[1]	Creating a Question Similarity mechanism that could identifies the unanswerable and irrelevant questions	BERT model, NLP	SQuAD 2.0 dataset, ProgphetNet	It helps the Question Answering Systems to focus only on the answerable questions to improve their performance. Introducing and application of the QAS that generates the question–answer pairs given a passage and is useful in several fields	Reading comprehension questions

Table 6 (continued)

No.	Refs	Purpose	Methods	Tools	Result	Type of Question
28	[14]	to generate the test questions that are mapped with bloom's taxonomy to tdetermine the learner's cognitive level	NLP, rule-based approaches	POSTag	The outputs are dynamic to create a different set of questions at each execution	Cloze type quetsions
29	[10]	Propse three novel algorithms to help educators generate questions to evaluate learners comprehension of the learning material	NLP, NLU, Machine comprehension	Stanford CoreNLP library, Python, Flask, IBM Cloud's Watson NLU service, Docker	Most of the questions generated by the AQG from the sample texts were valid. Invalid questions are likely the expected lower-scoring noun chunks	Multiple choice, fill-in-the-blank, true-or-false questions
30	[32]	Automatic Question Generation Model proposes a solution to save time, effort, and student's learning process which helps in educational purposes	Deep learning	SQuAD	This model gets a BLEU-4 score 11.3 which is good according to it generated automatically using deep learning approaches	Wh Questions
31	[33]	Present a method for automatic generation of affix based distractor for Tamil fill-in-the-blank question which are mainly used for learning Tamil grammar morphological details and vocabulary	NLP	ListNet, ListMLE, TamilMCQs	Overall, the method proposed pipelined process increases plausibility and reliability in distractor generation	fill-in-the-blank questions

the research conducted by Yuan et al. [42] produced questions of the same type as this study, namely the type of short answer questions. However, the evaluation carried out is different, in the research evaluation is carried out by comparing the system developed with the existing system in the previous study, namely "Seq2Seq" and the results using the Fluency (PPL) value. The value generated by the researcher is 175.7 while the "Seq2Seq" system has a value of 153.2. Which means that the system developed by the researcher produces more specific questions, although there are still shortcomings in terms of context.

5 Conclusions

After conducting research on automatic question generation with short answer questions on IELTS reading comprehension, the following are important contributions:

This research has succeeded in designing a computational model for automatic question generation with short answer questions on IELTS reading comprehension. This computational model uses the k-NN algorithm and the NLP method. The stages include scrapping to get news articles, tokenization, conversion to tree structure, simple sentence extraction, generating questions, cleaning question sentences that have pronouns, converting question sentences into POS tags, converting them to numeric, then calculating the distance between the test data. with training data using numeric data, changing words in sentences into synonyms, improving grammar, and finally removing question sentences from the same sentence or duplication.

After experimenting with five articles with five different themes and also four different websites, the results were in the form of candidate questions which were then evaluated by two experts. The results of the evaluation carried out showed that the grammatical correctness had a percentage of 59.52%, for answer existence it had a percentage of 95.24%, and for the difficulty index it had a percentage of 34.92%. So, it can be seen that the answer existence is very good.

Data availability All the data is collected from the simulation reports of the software and tools used by the authors. Authors are working on implementing the same using real world data with appropriate permissions.

Declarations

Competing interest Not Applicable.

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Aithal SG, Rao AB, Singh S (2021) Automatic question-answer pairs generation and question similarity mechanism in question answering system. Appl Intell 51(11):8484–8497
2. Ali, H., Chali, Y., and Hasan, S. A. (2010). Automation of question generation from sentences. In Proceedings of QG2010: The Third Workshop on Question Generation (pp. 58–67).
3. Alsubait, T., Parsia, B., and Sattler, U. (2012). Automatic generation of analogy questions for student assessment: an Ontology-based approach. Res Learn Technol. 20. vol. 20, pp. 95–101.

4. Blšták M, Rozinajová V (2022) Automatic question generation based on sentence structure analysis using machine learning approach. Nat Lang Eng 28(4):487–517
5. Bourne, P. (2005). High Impact IELTS Workbook Academic Module. Pearson.
6. Cambridge University (1996) Cambridge Practice Tests for IELTS 2. Cambridge University Press, New York
7. Cambridge University (2003) Cambridge IELTS 3 With, Answer. Cambridge University Press, New York
8. Chali, Y., & Hasan, S. A. (2012). Towards automatic topical question generation. In Proceedings of COLING 2012 (pp. 475–492).
9. Cotton K (1988) Classroom questioning School improvement research series 5:1–22
10. Cruz, R. R. D. L., Khalil, A., and Khalifa, S. (2021). Automatic multiple-choice and fill-in-the-blank question generation from arbitrary text. In Future of Information and Communication Conference (pp. 244–257). Springer, Cham.
11. Das B, Majumder M, Sekh AA, Phadikar S (2022) Automatic question generation and answer assessment for subjective examination. Cogn Syst Res 72:14–22
12. Fung, Y. C., Kwok, J. C. W., Lee, L. K., & Chui, K. T. (2020, August). Automatic question generation system for english reading comprehension. In International Conference on Technology in Education (pp. 136–146). Springer, Singapore.
13. Goto T, Kojiri T, Watanabe T, Iwata T, Yamada T (2010) Automatic generation system of multiple-choice cloze questions and its evaluation. Knowl. Manag. E-Learn: Int J 2(3):210–224
14. Gnanasekaran, D., Kothandaraman, R., and Kaliyan, K. (2021). An Automatic Question Generation System Using Rule-Based Approach in Bloom's Taxonomy. Recent Adv Comput Sci Commun (Formerly: Recent Patents on Computer Science), 14(5), 1477–1487.
15. Harrison, R., Pathare, E., Pathare, G., and May, P. (2013). Headway Academic Skills: IELTS Study Skills Edition. Level 1 Student's Book. Oxford University Press.
16. Heilman, M., and Eskenazi, M. (2007). Application of automatic thesaurus extraction for computer generation of vocabulary questions. In Workshop on Speech and Language Technology in Education. pp. 65–68, 2007.
17. Huang, Y. T., Chen, M. C., & Sun, Y. S. (2012). Personalized automatic quiz generation based on proficiency level estimation. In 20th International Conference on Computers in Education (ICCE 2012).
18. Jakeman V, McDowell C (2008) New Insight Into IELTS Student's Book Pack. Cambridge University Press
19. Jerome, B., Van Campenhout, R., & Johnson, B. G. (2021). Automatic question generation and the SmartStart application. In Proceedings of the Eighth ACM Conference on Learning@ Scale (pp. 365–366).
20. Kalady, S., Elikkottil, A., and Das, R. (2010). Natural language question generation using syntax and keywords. In Proceedings of QG2010: The Third Workshop on Question Generation (Vol. 2, pp. 5–14). questiongeneration. org.
21. Kim, M. K., & Kim, H. J. (2008). Design of question answering system with automated question generation. In 2008 Fourth International Conference on Networked Computing and Advanced Information Management (Vol. 2, pp. 365–368). IEEE.
22. Kumar, G., Banchs, R. E., and D'Haro, L. F. (2015). Revup: Automatic gap-fill question generation from educational texts. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 154–161).
23. Kusuma SF, Siahaan DO, Fatichah C (2022) Automatic question generation with various difficulty levels based on knowledge ontology using a query template. Knowl-Based Syst 249:108906
24. Liu M, Calvo RA, Rus V (2012) G-Asks: An intelligent automatic question generation system for academic writing support. Dialogue & Discourse 3(2):101–124
25. Lougheed, L. (2013). Barron's IELTS Practice Exams: With Audio CDs. Barron's.
26. Lougheed, L. (2016). Barron's IELTS: With Audio Cd. Barron's Educational Series.
27. Majumder M, Saha SK (2014) Automatic selection of informative sentences: The sentences that can generate multiple choice questions. Knowl. Manag. E-Learn: Int J 6(4):377–391
28. Mazidi, K., and Nielsen, R. (2014). Linguistic considerations in automatic question generation. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 321–326).
29. Mazidi, K., and Tarau, P. (2016). Infusing nlu into automatic question generation. In Proceedings of the 9th International Natural Language Generation conference (pp. 51–60).
30. McCarter, S. and Ash, J. (2001). IELTS Reading Tests, Future Publications by IntelliGene.
31. McDowell C, Jakeman V (1996) Cambridge practice tests for IELTS 1. Cambridge University Press

32. Mokhtar M, Doma S, Abdel-Galil H (2021) Automatic Question Generation Model Based on Deep Learning Approach. Int J Intelligent Com Inf Sci 21(2):110–123
33. Murugan S, Ramakrishnan BS (2022) Automatic Morpheme-based Distractors Generation for Fill-in-the-Blank Questions using Listwise Learning-To-Rank Method for Agglutinative Language. Eng Sci Technol An Int J 26:100993
34. Nguyen, H. A., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards Generalized Methods for Automatic Question Generation in Educational Domains. In European Conference on Technology Enhanced Learning (pp. 272–284). Springer, Cham.
35. Patrick H (2001) Master IELTS 6: IELTS Precise Reading. Xi An Jiaotong University Press, China
36. Riza LS, Pertiwi AD, Rahman EF, Munir M, Abdullah CU (2019) Question Generator System of Sentence Completion in TOEFL Using NLP and K-Nearest Neighbor. Indones J Sci Technol 4(2):294–311
37. Riza LS, Anwar FS, Rahman EF, Abdullah CU, Nazir S (2020) Natural Language Processing and Levenshtein Distance for Generating Error Identification Typed Questions on TOEFL. J Comput Soc 1(1):1–23
38. Shuai, P., Li, L., Liu, S., and Shen, J. (2022). QDG: A unified model for automatic question-distractor pairs generation. Applied Intelligence, 1–11.
39. Susanti Y, Tokunaga T, Nishikawa H (2020) Integrating automatic question generation with computerised adaptive test. Res Pract Technol Enhanc Learn 15(1):1–22
40. Tsumori S, Kaijiri K (2006) System Design for Automatic Generation of Multiple-Choice. Eng Educ 14(2):151–159
41. Van Geyte E, Snelling R (2011) Reading for IELTS. HarperCollins
42. Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Subramanian, S., and Trischler, A. (2017). Machine comprehension by text-to-text neural question generation. arXiv preprint arXiv:1705.02012., pp. 1–14.
43. Zhang X, Yan X, Yao Z (2021) The Automatic Question Generation System for CET. J Com Comm 9(9):161–168
44. Žitko B, Ljubić H (2021) Automatic question generation using semantic role labeling for morphologically rich languages. Tehnički vjesnik 28(3):739–745