Journal of Computers for Society
Journal homepage: https://ejournal.upi.edu/index.php/JCS
Vol. 1 No. 1, June-2020, pp. 1-23

Natural Language Processing and Levenshtein Distance for Generating Error Identification Typed Questions on TOEFL

Lala Septem Riza[a], Faisal Syaiful Anwar[b], Eka Fitrajaya Rahman[c], Cep Ubad Abdullah[d], Shah Nazir[e]

[a. b. c] *Department of Computer Science Education, Universitas Pendidikan Indonesia, Jl. Setiabudhi 229, Bandung, Indonesia*
[a]lala.s.riza@upi.edu, [b]faisalsyfl@gmail.com, [c]ekafitrajaya@upi.edu
[d] *Department of English Education, Universitas Pendidikan Indonesia, Jl. Setiabudhi 229, Bandung, Indonesia*
[d]cepubad@upi.edu
[e] *Department of Computer Science, University of Swabi, Swabi, Pakistan*
[e]snshahnzr@gmail.com

Abstract

Test of English as a Foreign Language (TOEFL) is one of the evaluations requiring good quality of the questions so that they can reflect the English abilities of the test takers. However, it cannot be denied that making such questions with good quality is time consuming. In fact, the use of computer technology is able to reduce the time spent in making such questions. This study, therefore, develops a model to generate error identification typed questions automatically from news articles. Questions from the sentences on news sites are created by utilizing Natural Language Processing, Levenshtein Distance, and Heuristics. This model consists of several stages: (1) data collection; (2) preprocessing; (3) part of speech (POS) tagging; (4) POS similarity; (5) choosing question candidates based on ranking; (6) determining underline and heuristics; (7) determining a distractor. Testing ten different news articles from various websites, the system has produced some error identification typed questions. The main contributions of this study are that (i) it can be used as an alternative tool for generating error identification typed questions on TOEFL from news articles; (ii) it can generate many questions easily and automatically; and (iii) the question quality are maintained as historical questions of TOEFL.

Keywords: TOEFL, Automatic Question Generation, Natural Language Processing, Levenshtein Distance

First draft received: May 27, 2020
Date Accepted: June 14, 2020
Final proof received: June 18, 2020

1. Introduction

One of the evaluations that requires high quality of questions is the Test of English as Foreign Language (TOEFL). It is a form of standardized English test destined for those who are not native speakers. It becomes one of the most taken evaluation because it is a prerequisite in the selection of new employee recruitment, the prerequisite of studying on master/ doctoral programs in domestic and international levels, even as the requirement of undergraduate candidates in several public and private universities in Indonesia [1]. There are several sections tested in the TOEFL [2], i.e., (i) listening comprehension consisting of short conversations, longer conversations, and lectures or talks, (ii) structure and written expression consisting of sentence completion and error identification, and (iii) reading comprehension. Moreover, since TOEFL is held almost all over the world and conducted in high frequencies, the questions of TOEFL need to be produced at high speed and in large quantities at all times while still maintaining its quality.

At this time, it cannot be denied that making the evaluation question is really time consuming for the test makers. Question makes usually spend about 20% -50% of their time thinking about one set of questions [3]. The use of computer technology is able to reduce the time spent by the question makers in creating the test questions [4]. With the help of technology, certain problems in education can be overcome. One of the technological sciences that can help is namely Natural Language Processing (NLP).

In recent years, the automated questioning system of a sentence has received more attention from NLP researchers [5]. Many researchers have been doing such a study to get high accuracy with various algorithms. One of the main objectives of this research is to extract keywords from a text to be converted into a question. For example, automated problem-generating systems can help the question makers in this grueling task, saving time and resources [6]. The research on Question Generation (QG), especially on making of 5W + 1H [5], and the most widely used question because it is most effective for honing students' knowledge of multiple choice [7].

In this research, we focus on generating error identification typed questions that are produced from news articles by utilizing NLP, Levenshtein Distance [8], and heuristics. Basically, the proposed model consists of several stages: (1) data collection; (2)

preprocessing; (3) part of speech (POS) tagging; (4) POS similarity; (5) choosing question candidates based on ranking; (6) determining underline and heuristics; (7) determining a distractor of the answer. The main advantages of this approach are that questions can be automatically generated in high numbers at the same time, and they have up to date contents. Moreover, the quality can be maintained by Levenshtein distance between candidate questions and historical TOEFL questions and determined heuristics. In addition, this study will evaluate the generated questions according to several parameters, namely grammatical correctness (GC), answer existence (AE), distractor quality (DQ) and Difficulty Index (DI). Thus, the quality of questions generated can be measured and analyzed.e not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

2. Methods

2.1 Natural Language Processing and Its Implementations on Generating Questions

Language as an important part of human life, in written form can be a record of the knowledge gained by mankind from one generation to the next while in the oral form is a means of communication between individuals in a society [9]. Therefore, the goal in the field of NLP is to make the process of computing model of the language, so that there can be an interaction between humans and computers with natural language mediation. This computational model can be useful for scientific purposes such as researching the properties of a natural language form as well as for everyday purposes in this case facilitate communication between man and computer.

For developing NLP, we should pay attention to the knowledge of the language itself, both in terms of the words used, how the words are combined to produce a sentence, what a word means, what is the function of a word in a sentence and so on. Moreover, we must also consider that there is one more thing that plays a significant role in language, that is, the human capacity to understand and the ability to obtain from the knowledge gained continuously during life. For example, in a conversation, a person may be able to answer a question or participate in a conversation not only based on language skills but also to know for example the term commonly used in the conversation group or even to know the context of the conversation itself.

NLP is a large area, covering topics such as text understanding and machine learning. One focus of NLP is information extraction, which processes text content so it can be incorporated into a relational database or analyzed using data mining. In the extraction of information, the text content is an insert. Whereas, the output is a data format defined in accordance with the required application. The information extraction system can be used to process large amounts of information, so adequate computer performance is required. Basically, there are five general stages of information extraction [10], including:

- Tokenizer: it is a process for dividing texts in forms of sentences, paragraphs or documents, into tokens / specific parts [11].
- Part of Speech Tagger: Part of speech is the parts used to form a sentence in the English language. In English, there are 8 kinds of Part of Speech covering verb, noun, pronoun, adjective, adverb, conjunction, preposition, and article. Part of Speech Tagging is a process of automatically labeling word classes in a word in a sentence [12].
- REGEX Matcher: Regex is a sequence of characters that determines search patterns. Regex is used to search, edit and manipulate texts. Regex has become the standard spread across all tools and programming languages so it is important to learn. Regex is one of the stages of information extraction because it can help the extraction of metacharacter, text, and other important parts. Regex utilization in technology that is search engines, search and replacement of dialog in word processing application and text editors, research on text processing and lexical analysis
- Filler & Merger Templates: Template filling is one of the efficient approaches for extracting a complex structure of information in texts. Template filling is an important role in Information Extraction (IE) and

Text Mining to unite information across multiple sentences to identify all the roles that are searched.

The automated questionnaire system is basically a technology that can help problem-makers (question makers) to facilitate problem-solving. Some research on automatic generate system has been done. A survey of related research studies on question generation is presented in **Table 1**. It is basically divided based on 3 sections, namely question type, question language, and methodology. According to question types, researchers have proposed a short field, 5W questions (What, when, where, who, why), and multiple choice. For the methodology used, we have named entity relationship, information classification, and knowledge descriptor. So far, the languages that can be generated in research question generation are English and Punjabi.

Table 1. *A short survey on implementations of NLP for generating questions*

Refs	Methodology	Question Type	Language
[13]	Sentence Classification	Who, what, when.	English
[14]	Sentence selection, NER	Cloze fill	English
[15]	Sentence selection, question construction	Gasp fill	English
[16]	Extract person, location, date	Where, who and when	Punjabi
[17]	Ontology-based strategies like class based, property based, terminology based strategies	Multiple choice	English
[18]	Document Processing, Information Classification and Question Generation.	Definition, description, example, and essay	English

2.2 Model Construction for Generating Error Identification Typed Questions

The proposed model for generating questions with the error identification type consists of several stages as follows (1) data collection; (2) preprocessing; (3) part of speech (POS) tagging; (4) POS similarity; (5) choosing question candidates based on ranking; (6) determining underline and heuristics; (7) determining a distractor of the answer as illustrated in **Figure 1**. Detailed explanations of each step in the figure are presented in the following subsection.

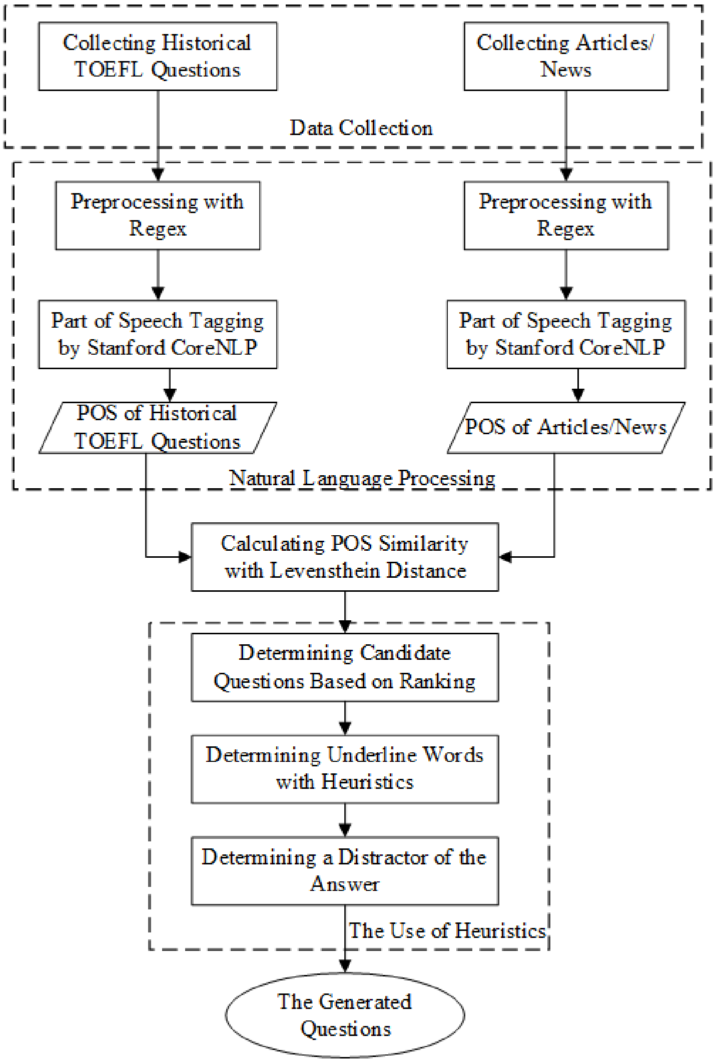
Collecting Historical TOEFL Questions
Collecting Articles/ News
Data Collection
Preprocessing with Regex
Preprocessing with Regex
Part of Speech Tagging by Stanford CoreNLP
Part of Speech Tagging by Stanford CoreNLP
POS of Historical TOEFL Questions
POS of Articles/News
Natural Language Processing
Calculating POS Similarity with Levensthein Distance
Determining Candidate Questions Based on Ranking
Determining Underline Words with Heuristics
Determining a Distractor of the Answer
The Use of Heuristics
The Generated Questions

Figure 1. *The research model for generating error identification typed questions.*

2.2.1 Data Collection for Model Construction and Question

Two data used are TOEFL's questions with the type of the error identification that are obtained from some textbooks and news/articles from public websites which are believed to have high grammar accuracy. Fort the first data, we collect the following datasets from the 8 following books, namely: TOEFL Test Preparation Kit [2];

Cambridge Preparation for the TOEFL Test [19]; Barron's How to prepare for the TOEFL [20]; CliffsTestPre: TOEFL CBT [21]; TOEFL Exam success from Learning Express [22]; TOEFL Grammar Review [23]; ETS TOEFL Practice Tests [2]; Longman Complete Course for the TOEFL Test [24].

The questions used as the datasets are limited to cover only the questions of the error identification type on each book. 3024 items of questions are taken and entered into the database, while the second dataset is a news article. Here are the news media sites used in this research: Al Arabiya (http://english.alarabiya.net); Al Jazeera (http://aljazeera.com); Australian Broadcasting Corporation (http://abc.net.au); Australian Broadcasting Corporation News (http://abcnews.go.com); BBC News (http://bbc.co.uk); CNN (http://cnn.com); Forbes (http://forbes.com); The Jakarta Post (http://thejakartapost.com); The Times (http://thetimes.co.uk); and The New York Times (http://nytimes.com).

2.2.2 Preprocessing on Data

After questions from the TOEFL books are collected on the previous step, the datasets need to be preprocessed, one of which is removing the punctuations because they contain many punctuationsthat cannot be recognized by POS Tagging. The preprocessing is done by utilizing one of the texts processing of regular expression (Regex). For example, we have here the following complete question with the answer: "The rain forest, with its large trees that provide shade to the vegetation below, is home to unique flora and fauna." The result after the Regex is "The rain forest with its large trees that provide shade to the vegetation below is home to unique flora and fauna". It can be seen that after Regex the sentence become the new one without punctuation. Thus, now the datasets can be processed at the next stage.

2.2.3 Part of Speech (POS) Tagging by Stanford CoreNLP

At this stage, each line in the question data will be processed through POS tagging to recognize the type of words grouped by function rather than each of the words [12].

In this study, we use one of the development libraries in computational linguistic, namely Stanford CoreNLP [11] accessible on https://stanfordnlp.github.io/CoreNLP/. There are 8 kinds of Part of Speech used, namely verb, noun, pronoun, adjective, adverb, conjunction, preposition, and article. In the known Tagging POS called tagset, the tagset is an English word class classification in the form of a tag. An example of a very popular tagset used is the Penn Treebank tagset [25]. Thus, the output of this stage will convert the English sentence into a Part of Speech sequence with the tagset. For example, the following sentence: "The rain forest with its large trees that provide shade to the vegetation below is home to unique flora and fauna" is changed into the following the part of speech tag: "DT NN NN IN VBZ JJ NNS WDT VBP NN TO DT NN IN VBZ NN TO JJ NN CC NNS". Then, we just repeat the same process for all sentences.

2.2.4 Calculating POS Similarity

After POS tagging on 2 datasets (historical TOEFL questions and news articles), we can compare the proximity between the datasets using POS similarity using Levenshtein Distance [8]. In other words, every POS tagging in news articles will be compared to all POS tagging in the TOEFL historical stories. The smaller value of the Levenshtein distance results indicates that two more data have similarities. As an illustration to provide a more detailed explanation, **Table 2** shows some distance results from one candidate question in an article against 3024 datasets in the historical TOEFL questions. It should be noted that POS tagging of news articles is called a candidate question. It can be seen in Table 2 that the smallest value in this scenario is 9. Therefore, the model will select and store the data index selected to enter into the next stage of ranking. We repeat the same processes for other candidates to get their distances.

2.2.5 Determining Candidate Questions Based on Ranking

Based on the results in the previous stage, we obtain a list of candidate questions on each article. The list is then sorted by the smallest distance that has been calculated in the previous stage. If the user wants to generate 5 questions in each article then we only take 5 candidates with the smallest distance in each article.

Table 2. *Results on calculating distance between candidate question and historical TOEFL questions.*

Candidate Question ID: 1		
POS Tagging of the question candidate: PRP VBD DT NN MD VB JJ NNS IN NN CC NN		
Id	**POS**	**distance**
1	WRB RB VBP PRP VB DT NN WDT VBD NNP PRP IN PRP$ NN NN	10
2	WRB PRP VBP DT NN WRB VBP PRP RB VB IN PRP IN NN IN NNP PRP VBP NN TO DT NN PRP MD VB PRP WP TO VB PRP VBD IN PRP$ NNS VBN IN DT NN VBD IN DT NN POS NNP NN	36
3	VBZ DT NNS POS NN VBN CC RB JJ IN NNS VBP CC NNS VBP VBG NN MD VB NNS CC NN CC PRP VBZ RB RB IN DT NN	24
4	VBN IN NNP WRB PRP MD VB IN DT NNP DT NN VBD EX MD VB DT JJ NN	15
5	VBN CC VBN PRP VBD VBN IN DT NN CC VBN TO NNP IN DT JJ NN IN PRP$ NN CC NN	13
6	PRP VBD PRP WP VBD VBG IN DT NN IN DT NN	9
7	VBG IN NN RBR NNS CD NNP VBP RB VBN RB CC IN JJ NNS VBD IN DT VBG NNS CC NNS	18
8	VB VBZ NN IN JJ NN VBZ IN PRP IN DT NN NN VBD PRP$ NN CC PRP$ NN TO DT NNS	18
9	VB IN JJ NN TO VB NN NNS CD NNS IN DT NN VBZ RB VBN VBN IN DT NN	14
10	RBR DT NN DT NNP NNP VBD DT NN NNS TO DT DT NN DT NN NN IN NN IN DT NN TO VB CD IN NN IN NNS VBD VBN	25
...		
3024	DT NN VBP DT TO JJ IN PRP TO VB IN DT NN CC RB PRP MD VB RB	15

2.2.6 Determining Underline Words with Heuristics (Frequent Tag)

It is the stage where selection of options or underline is carried out. A TOEFL problem with error type identification has the characteristics of having 4 underlined options. These options will be automatically generated by the system using the index selected in the POS similarity stage. For example, the index of selected training data has 4 underlines of PRP, WP, IN, and DT. Thus, the underline on the new candidate will follow the underline of the training data PRP, WP, IN, DT. After having four selected underlines, one of the four underlines will be a distractor or incorrect grammatical word. We choose this word by generating a random number.

2.2.7 Determining a Distractor of the Answer

After going through the POS similarity & ranking, selection of options (underline determination) and generation of random values, then the final stage in generating a question is determining the distractor (i.e., a word with a grammatical error to be the answer). Any problem that already has 4 options specified in the previous stage, the option is still a true vocabulary. Therefore, it is necessary to select one of the options, then the selection of diversity in the option. Thus, the preceding vocabulary is wrong.

The choice of the distractor has several rules to apply to the model, the rule is made so that the selection of the decoy does not make the problem to be ridiculous or too easy to work on. Here are the rules that the authors created:

1. Distractors for the verb words with the following POS tagging: VB, VBD, VBG, VBN, VBP, dan VBZ: The verb selection of verbs is taken from an online English dictionary using the Application Programming Interface (API). The API used is Ultralingua API on http://api.ultralingua.com/ page. The feature used in the API is the verb conjugation, where the API will generate all the equivalents of a similar word from the verb. For example:
    - Before: When goshawk chicks are young, both parents share in the hunting duties and in guarding the nest.

- In case we choose the verb *share* become the answer or incorrect grammatical word, then according to the heuristic, we change *share* into *had shared* to be the distractor.
- After: When goshawk chicks are young, both parents had shared in the hunting duties and in guarding the nest.

2. Distractors for the preposition words with the following POS tagging: IN and TO: If the answer / tag selected is the preposition (IN and TO), then we will take the choice of distractor by heuristics. Since we define that the preposition words consists of the following words: aboard, about, above, over, after, against, beside, along, behind, beside, besides, below, beneath, between, except, for, from, in, into, like, from, on, since, till, with and wthout, the choice of distractor is one of those words besides the word itself.
3. Distractors for the pronoun words with the following POS tagging: PRP and PRP$: If the selected tag is PRP or PRP $, then the specter will be selected according to use in subjective, objective, and possesive forms. For example if the choice is he then the exact observer is him and his:
   - Before: He also mentioned he would be speaking to the country's President Michel Aoun,
   - If we change the word he to his according to our heuristics.
   - After: His also mentioned he would be speaking to the country's President Michel Aoun,
4. Distractors for the modal words with the following POS tagging: MD: The most appropriate use of the cursor for modals (MD) is the past word of the word itself. For example if the correct answer is the word "*can*" then the distractor is "*could*".
5. Distractors for the determiner words with the following POS tagging DT: The choice of pata on the determiner (DT) only focuses on "a", "an" and "the". The decoy will be selected in accordance with the a, an and the options and in addition to the word itself. For example,
   - Before: Tell the Santa story in a way that connects to the Christmas story.
   - If the distractor for determiner the to be changed by an.

- After: Tell an Santa story in a way that connects to the Christmas story.

2.3 Experimental Design

In the experiments, we perform the system that implements the proposed model for generating TOEFL questions in the type of error identification. In this scenario, the system will generate 50 error identification problems from 10 different news articles. In other words, each article will produce each of the 5 questions, so that the total number of questions is 50. All the resulting questions will have 4 underlines including one word as the answer, which is the word with incorrect grammar. As we mentioned previously, we choose 10 URL of news websites with different topics of articles as illustrated in **Table 3**.

Table 3. *Datasets for testing in generating error identification typed question.*

No	Filenames	News URL	Topics	Numbers of Generated Questions
1	abcau.txt	http://abc.net.au	Politics	5
2	abcnews.txt	http://abcnews.go.com	Politics	5
3	alarabiya.txt	http://english.alarabiya.net	Security and Defense	5
4	aljazeera.txt	http://aljazeera.com	Security and Defense	5
5	bbc.txt	http://bbc.co.uk/news	Historics	5
6	cnn.txt	http://cnn.com	Entertaintment	5
7	forbes.txt	http://forbes.com	Current News	5
8	thejakartapost.txt	http://thejakartapost.com	Sport	5
9	thetimes.txt	http://thetimes.co.uk	Current News	5
10	theny.txt	http://nytimes.com	Government	5
			Total:	50

After running the experiments, we perform three aspects of analysis as follows:

1. Analysis with Grammar Checker: This analysis will prove whether the resulting question is also declared wrong by the grammar checker. The website used in this check is accessible on the www.nounplus.net/grammarcheck page.
2. Analysis on distractor by human experts: This analysis will prove whether the problem and the key answers generated by the system, stated according to the expert. Each expert will be presented 50 questions from experiments without key answer. All questions answered by the expert will be matched with the answer key. If appropriate it will be symbolized by the number 1, otherwise if one will be marked with number 0.
3. Evaluation and analysis on the question quality by human experts: Once the questions are generated, then there are stages in which the matter will be evaluated by the human experts. It is important to ensure the quality of the questions generated; otherwise, the problem cannot be used for the intended purpose. Therefore, we adopted a metric for evaluation proposed by [26] as follows:
   a. Grammatical Correctness (GC): It determines whether a question is syntactically well formed. The author determines 3 points to show the matrix scale based on the number of grammar error. Grammar error is calculated in addition to a clue that makes the sentence wrong:
      - 1: (best): Question does not have grammatical errors.
      - 2: questions have 1 or 2 grammatical errors.
      - 3 (worst): Questions have 3 or more grammatical errors.

   b. Distractor Quality (DQ): It is an assessment to measure how precisely a pervert of the four underlines is raised. The author makes a two-point scale for this assessment as follows:
      - 1 (worst): Distractor can be easily identified as wrong answers
      - 2 (best): Distractor can be feasible.

   c. Difficulty Index (DI): It is an assessment of how difficult the question generated from the system. This assessment is determined by all aspects of both questions and checkers. The author makes a scale of 3 points as follows:

- 1 (easy): The generated question is considered easy.
- 2 (medium): The resulting question is considered sufficient.
- 3 (hard): The resulting question is considered very difficult.

3. Results and Discussion

3.1 Experimental Results

This result is 50 error identification questions along with key answers. All these questions will then proceed to expert judgment to evaluate the quality of the questions. **Table 4** shows some questions generated by the system in this experiment.

Table 4. *Results on the experiments.*

No	Generated Questions	(Index) Answer
1	.Mural attacks escalate a same mural was targeted by another man on Friday	(1) the
2	Mr Morrison agreed the bill needed to address very fundamental issues besides faith and belief.	(4) of
3	It listed Mr Berry as the one needing protection due to “ongoing issued” with his neighbour.	(4) issues
…	…	…
50	Israeli Culture Minister Miri Regev say she hoped Lorde would reconsider her decision.	(2) said

3.2 Discussion

On this section, we analyze the results according to the experimental design on the previous section, namely analysis with Grammar Checker, analysis on distractor by human experts, and evaluation and analysis on the question quality by human experts.

3.2.1 Analysis with Grammar Checker

As we mentioned, this analysis will prove whether the resulting problem is declared wrong by the grammar checker. The website used in this check is accessible on the www.nounplus.net/grammarcheck page. Errors were happened because of grammatical error, misspelling, uncertainty, and undefined meanings. According to the experiments, there are only 23 out of a total of 50 questions which are declared wrong. Thus, the percentage of quality questions by grammar checker is 46%. However, mostly the inaccurate questions occur when the distractor/answer is on the preposition word.

3.2.2 Evaluation and analysis on the question quality by human experts

According to the experimental design, we evaluate the question quality are determined by four aspects: grammatical correctness (GC), answer existence (AE), distractor quality (DQ), and difficulty index (DI). It can been from Table 5 that the averages of two human experts on grammatical correctness, answer existence, distractor quality, and difficulty index are 1.08, 1.06, 1.66, and 1.57, respectively. Furthermore, the calculation using rating scale method will be categorized five categories by using the scale as follows: very good (i.e., between [80%, 100%]), good (i.e., between [60%, 80%]), enough (i.e., between [40%, 60%]), bad (i.e., between [20%, 40%]), and very bad (i.e., less than 20%). So, the question quality can be presented as **Table 5**.

3.2.3 Comparison with other researches

This section attempts to compare the proposed model and its implementation in this research with previous and relevant researches. The detailed comparison can be seen in **Table 6**. It can be seen that it is only this research that focuses on error identification typed question on TOEFL. Moreover, there is no a system, except this research, that maintain the quality of questions by considering historical TOEFL's questions as data training, even though there are several frameworks that can be used for TOEFL questions (e.g., filling in the blank, reading comprehension, etc).

Table 5. *Analysis on the question quality.*

Parameters	Ideal values	∑ Score per Parameter	Percentage	Categories
Grammatical Correctness	1	1.06	94%	Very good
Answer Existence	1	1.08	92%	Very good
Distractor Quality	2	1.72	86%	Very good
Difficulty Index	3	1.69	56%	Enough
		Average	82%	

4. Conclusion

After doing research on the implementation of Natural Language Processing, Levenshtein Distance, and heuristics on generating the error-identification typed questions, we can draw the following conclusions:

1. This research succeeded in making a model and its implementation of error-identification typed questions for TOEFL using the following steps: (1) data collection; (2) preprocessing; (3) part of speech (POS) tagging; (4) POS similarity; (5) choosing question candidates based on ranking; (6) determining underline and heuristics; (7) determining a distractor of the answer.
2. The questions generated are analyzed with several aspects, such as analysis with Grammar Checker, analysis on distractor by human experts, and evaluation and analysis on the question quality by human experts.

According to the results and their analysis, we can state that main contributions are that it can be used as an alternative tool for generating error identification typed questions on TOEFL from news articles easily and automatically while the quality of generated questions are maintained as historical questions of TOEFL.

For the future work, we have a plan to extend the model for sentence completion and reading comprehension in TOEFL. Other methods involving machine learning can be considered as well, such as methods based on fuzzy sets [27] and rough sets [28, 29].

Table 6. *Comparison with other systems.*

Ref	Methodology/Method/Algorithm	Input Data	Question Type	Language	Note
[13]	Syntactic parsing, Part Of Speech (POS) tagger and Named Entity analyzer.	Sentences provided by the Question Generation Shared Task Evaluation Challenge 2010	Who, what, when, where, why, and how many	English	A question is generated by 90 predefined rules expressing word interaction.
[14]	Three modules: sentence selection, keyword selection and distractor selection	English articles, i.e., Cricket World Cup 2011 data	Cloze/ fill-in-the-blank questions	English	Evaluation is done in three phases: selected sentences, selected keywords, and selected distractors.
[15]	Three stages are 1) sentence selection, 2) question construction, and 3) classification/scoring.	Articles, i.e., 105 articles from Wikipedia	Cloze/ fill-in-the-blank questions	English	Evaluation and analysis were conducted by ROC, error analysis, and feature analysis.
[17]	Several strategies based on ontology domain and knowledge base developed in Ontology Web Language (OWL)	Five ontologies from different domains	Multiple choice questions	English	The question has the same stem, which is “Choose the correct sentence:”
[18]	Four main steps: document processing agent, information classification agent, rules and template DB, and question generation	Text file	Multiple choice questions	English	Questions are generated by template provided in database. The templates are based on Bloom’s taxonomy.
[26]	The model contains several steps: question target selection, question construction, question template, distractor question, and multiple choice question construction.	Text: the ProcessBank corpus consisting of 200 paragraphs about biological processes, extracted from the high school level textbook Biology.	multiple-choice questions	English	Evaluation criteria consists of grammatical error, answer existence, inference step, and distractor quality.

Ref	Methodology/Method/Algorithm	Input Data	Question Type	Language	Note
[30]	Three steps: (1) extracting appropriate sentences for questions from texts based on Preference Learning, (2) estimating a blank part based on Conditional Random Field, and (3) generating distracters based on statistical patterns of existing questions	Articles and learning data from 1560 questions in TOEIC workbooks	Multiple choice cloze questions	English	Some evaluations have been used, such as ranking vote perceptron, 10 fold cross validation, and human experts.
[31]	The candidate are taken from the WordNet lexical dictionary for generating question options and filtered by Web searching.	-	English vocabulary test in reading comprehension of TOEFL: (1) a target word, (2) a reading passage in which the target word appears, (3) a correct answer, and (4) distractors (incorrect options)	English	This research adopted TOEFL vocabulary questions as the format. However, the steps involved in the model for generating questions are not so clear to be explained.
[32]	The model contains several steps: choosing target words (i.e., contextual scope and word co-occurrences), choosing distractors, and question generation.	Text, i.e., 1000 reading comprehension text passages obtained from ReadWorks.org	Fill-in-the-Blank questions	English	The questions were validated by 67 native English-speaking volunteers.
[33]	A template-based method which uses the structure of sentences to create multiple sentence patterns on various levels of abstraction	Text	What, where, which, how, and who questions	English	The model allows to create questions on different levels of difficulty and generality e.g. from general questions to specific ones.
[34]	The framework contains two main parts: strategic competence (i.e., sentence selection, paraphrasing, and question generation) and linguistic competence (i.e., semantic network, sematic role,	Text	Short answer questions for reading comprehension assessment	English	Many experiments have been conducted to measure performance of sentence selection module, synonym paraphrasing module, question generation module, the whole system, and post-edited items.

Ref	Methodology/Method/Algorithm	Input Data	Question Type	Language	Note
	latent semantic space, and lexical functional grammar)				
[35]	The system contains the following steps: preprocessing of input text, sentence selection, and key or blank word identification	History Books for School-Level Evaluation	Fill-in-the-blank questions	English	The validation was performed by 5 human evaluators.
This research	The computational model consists of : (1) data collection; (2) preprocessing; (3) part of speech (POS) tagging; (4) POS similarity; (5) choosing question candidates based on ranking; (6) determining underline and heuristics; (7) determining a distractor of the answer	News articles from websites	Error Identification Typed Questions on TOEFL	English	The analysis aspects contains analysis with Grammar Checker, analysis on distractor by human experts, and evaluation and analysis on the question quality by human experts.

References

[1] Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

[2] ETS. (2003). *TOEFL Practice TESTS volume 1*. Princeton.

[3] Cotton, K. (2001). *Classroom questioning. School improvement research series*. *3*.

[4] Aldabe, I., Lacalle, M. L. D., Maritxalar, M., Martinez, E., & Uria, L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 584–594. doi: 10.1007/11774303_58

[5] Agarwal, M., Shah, R., & Mannem, P. (2011). Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 1-9.

[6] Skalban, Y., Specia, L., & Mitkov, R. (2012). Automatic question generation in multimedia-based learning. In *Proceedings of COLING 2012: Posters*, 1151-1160.

[7] Hoshino, A., & Nakagawa, H. (2005, June). A real-time multiple-choice question generation for language testing: a preliminary study. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, 17-20.

[8] Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys (CSUR)*, *33*(1), 31–88. doi: 10.1145/375360.375365

[9] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[10] Riloff, E. (1999). Information extraction as a stepping stone toward story understanding. *Understanding language understanding: Computational models of reading*, 435-460.

[11] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

[12] Jurafsky, D. & Martin, J. H. (2014). S*peech and language processing*. Pearson Education India.

[13] Ali, H., Chali, Y., & Hasan, S. A. (2010). Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation*, 58-67.

[14] Narendra, A., Agarwal, M., & Shah, R. (2013). Automatic cloze-questions generation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 511-515.

[15] Becker, L., Basu, S., & Vanderwende, L. (2012). Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 742-751.

[16] Garg, P., & Bedi, E. S. (2014). Automatic Question Generation System from Punjabi Text using Hybrid approach. *International Journal of Computer Trends and Technology*, *21*(3), 130–133. doi: 10.14445/22312803/ijctt-v21p125

[17] Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic Generation of Multiple Choice Questions From Domain Ontologies. In *e-Learning*, 427-434.

[18] Pandey, S., & Rajeswari, K. C. (2013). Automatic Question Generation Using Software Agents for Technical Institutions. *International Journal of Advanced Computer Research*, *3*(13), 307-311.

[19] Gear, J., & Gear, R. (2002). *Cambridge Preparation for the TOEFL® Test Book with CD-ROM* (Vol. 1). Cambridge University Press.

[20] Sharpe, P. J. (2004). *How to Prepare for the TOEFL*. Univ of California Press.

[21] Pyle, M. A. (2000). *Cliffs Test Prep TOEFL CBT*. Foster City.

[22] Chesla, E. (2002). *TOEFL Exam success from LearningExpress*. New York: LearningExpress.

[23] Cemaiiiko, J. (2001), *TOEFL Grammar Review*, Rusia.

[24] Phillips, D. (2001). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests*. London: Longman.

[25] Marcus, M. P. & Santorini, B. (1993). *Building a large annotated corpus of English: The Penn treebank*. Computational Linguistics, vol. 19(2). 313-330.

[26] Araki, J., Rajagopal, D., Sankaranarayanan, S., Holm, S., Yamakawa, Y., & Mitamura, T. (2016). Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1125-1136.

[27] Riza, L. S., Bergmeir, C. N., Herrera Triguero, F., & Benítez Sánchez, J. M. (2015). frbs: Fuzzy rule-based systems for classification and regression in R. *Journal of Statistical Software, 65(6),* 1-30.

[28] Riza, L. S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślezak, D., & Benítez, J. M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package "RoughSets". *Information Sciences*, *287*, 68-89.

[29] Nazir, S., Shahzad, S., & Riza, L. S. (2017). Birthmark-based software classification using rough sets. *Arabian Journal for Science and Engineering*, *42*(2), 859-871.

[30] Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal*, *2*(3), 210-224.

[31] Susanti, Y., Iida, R., & Tokunaga, T. (2015). Automatic generation of english vocabulary tests. In *CSEDU (1)*, 77-87.

[32] Hill, J., & Simha, R. (2016). Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 23-30.

[33] Blšták, M., & Rozinajová, V. (2016). Automatic question generation based on analysis of sentence structure. In *International Conference on Text, Speech, and Dialogue,* 223-230.

[34] Huang, Y., & He, L. (2016). Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, *22*(3), 457-489.

[35] Pannu, S., Krishna, A., Kumari, S., Patra, R., & Saha, S. K. (2018). Automatic Generation of Fill-in-the-Blank Questions from History Books for School-Level Evaluation. In *Progress in Computing, Analytics and Networking*, 461-469.