Date of publication DEC-31, 2025, date of current version AUG-08, 2025
www.computingonline.net/ computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.4.4331

Indo-WDSimpleQuAD2.0: an Indonesian Benchmark Dataset for Knowledge Graph Question Answering System

MOHAMMAD YANI[1], WAWAN SETIAWAN[2], RIZKY FRIHATMAWATI[3], WALI ATMAMUDIN[4], MUHAMAD MUSTAMIIN[1], RENDI[1], ESTI MULYANI[1], FACHRUL PRALIENKA BANI MUHAMAD[1], ADILA ALFA KRISNADHI[5], INDRA BUDI[5]

[1]Politeknik Negeri Indramayu, Indramayu, Indonesia
[2]Universitas Singaperbangsa Karawang, Karawang, Indonesia
[3]Universitas Bina Sarana Informatika, Jakarta, Indonesia
[4]Excel Translation, Depok, Indonesia
[5]Universitas Indonesia, Depok, Indonesia

Corresponding author: Mohammad Yani (e-mail: mohammad.yani@polindra.ac.id).

ABSTRACT We propose Indo-WDSimpleQuAD2.0, a silver standard for an Indonesian-language benchmark dataset developed from SimpleQuestions and LC-QuAD 2.0 based on Wikidata. This dataset development is proposed due to the current absence of a representative KGQA benchmark dataset in Indonesian language. SimpleQuestions and LC-QuAD 2.0 were chosen because, in terms of question type variety and complexity, these datasets serve as supersets of other available datasets. Indo-WDSimpleQuAD2.0 comprises 27,924 questions for SimpleQuestions and 31,821 for LC-QuAD 2.0. Indo-WDSimpleQuAD2.0 was developed through a rigorous translation process by English language experts and native Indonesian speakers. This translation process was conducted in three rigorous stages: initial translation, validation and verification, and finalization of the translation. To ensure the quality of this dataset, the authors applied four criteria: translation accuracy, writing quality, semantic integrity, and annotation process. Indo-WDSimpleQuAD2.0 can serve as the first Indonesian-language KGQA benchmark dataset based on Wikidata, thus supporting future research and development of Indonesian KGQA systems.

KEYWORDS Indonesian benchmark; Indonesian dataset; KGQA; KGQA system evaluation.

I. INTRODUCTION

In recent years, research on Knowledge Graph Question Answering (KGQA) systems has advanced rapidly. KGQA systems are question-answering systems that utilize knowledge graphs (KG) as their data source. A KG is data modeled in the form of a set of triples, consisting of a subject, predicate, and object, and is expressed in a specific language known as the Resource Description Framework (RDF) [1]. A Question Answering (QA) system is a system in which the input is a question in natural language, and the output is an accurate answer to that question [2] [3]. Meanwhile, a KGQA system is a QA system that uses a knowledge graph (KG) as its data source [4]. In a Knowledge Graph Question Answering (KGQA) system, questions are decomposed into a set of entities and relations. These entities and relations are then structured into a query using a standardized language known as SPARQL to retrieve answers from the Knowledge Graph (KG) [5]. In the context of a KGQA system, the natural language input received by the KGQA system is subsequently translated into SPARQL format to retrieve data from the KG [6].

In general, two main approaches are employed in developing a Knowledge Graph Question Answering (KGQA) system: non-deep learning and deep learning. A portion of KGQA systems are developed using the non-deep learning approach [7]. Meanwhile, other KGQA systems are developed using the deep learning approach [8]. KGQA systems that utilize the deep learning approach require high-quality

datasets to achieve optimal results. These datasets serve not only as training data but also as test data to evaluate the performance of the developed KGQA system. The datasets used are benchmark datasets specifically designed for training KGQA models and assessing their performance. Currently, some KGQA systems also employ Large Language Models (LLMs) such as BERT to obtain answers that are relevant to increasingly complex questions [9].

Several benchmark datasets for KGQA systems are currently available and widely used, ranging from datasets for simple questions to those for complex queries [10]. Firstly, there is the QALD 1-9 dataset series. The QALD dataset series comprises 18 different datasets, each containing between 41 and 408 questions [10]. The second is the SimpleQuestions dataset. SimpleQuestions is a dataset containing over 108,000 questions in the form of simple queries [11]. In its development, SimpleQuestions is also available in a Wikidata version. The Wikidata version of SimpleQuestions comprises approximately 27,000 questions, as examined by researchers [12]. Next is SimpleDBPediaQA, a derivative dataset of SimpleQuestions based on DBpedia [13]. Another dataset is WebQuestions, which is based on the Freebase KG and contains 3,778 questions for training data and 2,032 questions for testing data [14]. The next dataset is LC-QuAD 2.0. This dataset is a large collection containing over 30,000 questions, which include both simple and complex queries [15].

Among the various benchmark datasets mentioned above, each has its own advantages and disadvantages. SimpleQuestions has a large amount of data but does not support multilingual data and focuses only on simple question types. QALD supports multilingual data but has a very limited amount of data, making it less ideal for training using machine learning approaches. LC-QuAD 2.0 supports multilingual data, has a large amount of data, and features more diverse question types, but it still remains very limited for simple question types. By developing Indonesian-language benchmark datasets for SimpleQuestions and LC-QuAD 2.0, we can address issues related to multilingualism, data quantity, and question diversity in benchmark datasets for KGQA systems simultaneously. This research aims to develop Indonesian-language benchmark datasets for KGQA systems for SimpleQuestions and LC-QuAD 2.0. The anticipated contribution of this research is to produce a benchmark dataset for Indonesian KGQA systems, thereby fostering the advancement of research in Indonesian-language KGQA systems.

II. RELATED WORK

Currently, there are several KGQA systems for both simple and complex questions. Usually, the developed KGQA systems focus on addressing issues in the tasks of the KGQA system, such as entity detection, entity prediction, relation prediction, answer matching, and subgraph selection. Research [16] utilizes the BERT model to solve issues related to entity detection. Studies [17] and [18] employ Bi-LSTM with attention for entity prediction. Meanwhile, for linking entities to the knowledge graph (KG), an approach using CNN with adaptive max pooling is used by [19]. Additionally, the use of CNN with adaptive max pooling in [19] can also predict relations present in the questions. Research [20] employs a custom architecture consisting of BERT, relation-aware attention networks, Bi-LSTM, and linear layers to match answers with data in the KG. There are also KGQA systems developed to find answers to given questions by traversing subgraphs in the KG, by performing n-gram matching between the text in the question and strings in the subgraph of the KG [19]. The key point here is that the availability of a representative benchmark dataset to solve several issues in the tasks of the KGQA system is highly crucial. This is a consideration that a good dataset can facilitate researchers in developing their studies related to KGQA systems. Among the various benchmark datasets listed below, there are five benchmark datasets that are most widely used.

A. SERIES QALD

The QALD dataset series consists of nine series. The emergence of the QALD series began with a KGQA system competition in 2011.[1] The competition presented challenges to KGQA system researchers, particularly concerning multilingualism [21]. This dataset is relatively small, containing only 41 to 408 questions. Given its limited size, this dataset is not suitable for developing KGQA models that employ a deep learning approach.

In addition to the multilingualism issue, the QALD-4 series [22], QALD-5 [23], and QALD-9 [24] also present challenges related to biomedical issues and hybrid QA systems that integrate Knowledge Graphs and conventional data. QALD-7 provides challenges concerning large-scale KGQA systems [25].

B. SIMPLEQUESTIONS

This dataset was published in 2015 by Border et al. [11]. It contains a collection of simple questions, with its Knowledge Graph based on Freebase. In its initial publication, the dataset comprised two variants: FB2M and FB5M. FB2M includes 2 million entities and 5,000 relations, while FB5M consists of 5 million entities and 7,500 relations. The SimpleQuestions Freebase dataset contains a total of 108,000 questions. Over time, SimpleQuestions has also become available in the Wikidata Knowledge Graph.[2]

C. SIMPLEDBPEDIAQA

This dataset was first introduced by Azmy et al. in 2018 [13]. It is a derivative of SimpleQuestions, where the questions from SimpleQuestions Freebase are mapped to DBpedia. SimpleDBPediaQA contains a total of 43,000 questions.

[1]https://github.com/ag-sc/QALD

[2]https://github.com/askplatypus/wikidata-simplequestions/tree/master/SimpleQuestions_v2

D. WEBQUESTIONS

This dataset also utilizes Freebase as its Knowledge Graph. It was published by Brown et al. [14] and contains 3,778 questions for training and 2,032 questions for testing. The dataset employs the JSON format. Its contents consist of three components: the URL from Freebase, the target value, and the expression (question).

E. LC-QUAD 2.0

This dataset contains 21,000 entities and 1,300 unique relations, along with 30,000 unique SPARQL queries. It includes 10 different variants of questions. Although this dataset is intended for complex question types, it also contains simple questions [15]. With such a wide variety of questions, this dataset can be considered a superset of the existing datasets for KGQA systems.

Among the five datasets presented above, none are available in the Indonesian language. This lack of benchmark datasets for Indonesian-language KGQA systems poses a significant issue, as it may restrict the development of KGQA systems in Indonesian.

III. INDO-WDSIMPLEQUAD2.0

Indo-WDSimpleQuAD2.0 is a dataset developed from the SimpleQuestions and LC-QuAD 2.0 datasets based on Wikidata, specifically tailored for the Indonesian language. Indo-WDSimpleQuAD2.0 comprises two datasets: Indonesian-language SimpleQuestions and Indonesian-language LC-QuAD 2.0. Indo-WDSimpleQuAD 2.0 can be accessed at https://github.com/moh-yani/indo-wdsimplequad20.

The Indonesian-language SimpleQuestions dataset consists of three files: training data, validation data, and testing data, containing 19,481, 2,821, and 5,622 questions, respectively, for a total of 27,924 questions. This dataset file is in CSV format. Each file in this dataset comprises four columns. The first three columns represent triples (in ID format) as answers to the questions, while the fourth column contains the questions themselves. In each file, odd-numbered rows contain questions in English along with their corresponding answer triples, whereas even-numbered rows feature questions in Indonesian alongside their answer triples. Table 1 presents the metadata for the SimpleQuestions section of Indo-WDSimpleQuAD2.0.

Table 1. Metadata for Indo-WDSimpleQuAD2.0: SimpleQuestions Section

Row	Column 1 to 3	Column 4
Odd	Triple ID (subject, predicate, object)	Question in English
Even	Triple ID (subject, predicate, object)	Question in Indonesian

As shown in Table 1, a triple refers to a statement represented in the form of subject, predicate, and object. For example, a statement from the SimpleQuestions test data, "*Roger Marquis meninggal di Holyoke*", is represented as the triple ID (Q7358590, P20, Q1637790). The triple ID corresponds to the Wikidata IDs of the entities and relations. Entity IDs begin with the letter "Q", while relation IDs begin with the letter "P". Columns 1-3 represent the answer pairs, and column 4 contains the questions. In the example statement "*Roger Marquis meninggal di Holyoke*", this fact serves as the answer to the question "*Di mana Roger Marquis meninggal?*".

The Indonesian-language LC-QuAD 2.0 dataset consists of two files: training and testing data, containing 25,438 and 6,383 questions, respectively. This dataset file is in JSON format. Each question comprises 12 fields, which are: NNQT_question, uid, subgraph, template_index, question, question_ina, sparql_wikidata, sparql_dbpedia18, template, answer, template_id, and paraphrased_question. The questions in Indonesian are found in the field question_ina. Table 2 provides an explanation of the metadata structure for the LC-QuAD 2.0 section of the Indo-WDSimpleQuAD2.0 dataset.

Table 2. Metadata for Indo-WDSimpleQuAD2.0: LC-QuAD 2.0 Section

Field	Description	Data Type
NNQT_question	Question generated by the system	String
uid	Unique ID number	int32
subgraph	Subgraph of question	String
template_index	Index of question template	int32
question	Verbalized question	String
question_ina	Indonesian verbalized question	String
sparql_wikidata	SPARQL query from Wikidata endpoint	String
sparql_dbpedia18	SPARQL query from DBPedia endpoint	String
template	Question template	String
answer	The answer	String
template_id	Template ID	int32
paraphrased_question	Paraphrased question	String

Essentially, the questions in the "question" field are verbalizations of the original questions (NNQT) generated from the SPARQL query syntax to obtain answers to the questions.

IV. METHODOLOGY

In this section, the author will explain how Indo-WDSimpleQuAD2.0 was developed. The choice of SimpleQuestions and LC-QuAD 2.0 based on Wikidata as the foundational datasets for this research is due to two main reasons. First, the SimpleQuestions and LC-QuAD 2.0 datasets encompass a variety of question types and challenges that reflect the variations found in other datasets. Second, the Wikidata Knowledge Graph remains active and has been rapidly evolving to date [26], in contrast to Freebase, which was shut down and its access and development ceased as of August 31, 2016. Additionally, Wikidata supports cross-lingual needs [27], thereby allowing for the potential development of multilingual KGQA systems based on this knowledge graph. Moreover, Wikidata can also be

utilized for sharing and exchanging metadata repositories [28].

A. CURATION AND TRANSLATION

The datasets used in this research are derived from LC-QuAD 2.0 and SimpleQuestions based on Wikidata.[3] To achieve high-quality translations, we employed a direct human translation approach. The translation and validation processes were conducted by experts in English and native speakers of Indonesian.

To ensure the quality of the Indo-WDSimpleQuAD2.0 dataset, we employed four criteria: translation accuracy, writing quality, semantic integrity, and annotation process. The criteria for writing quality and the annotation process were adapted from [29]. The annotation process refers to the verification of entities and relations present in the questions. The accuracy of the translations was checked manually by experts in English and native speakers of Indonesian. Manual checks were also conducted to verify the quality of the writing by reviewing each translated item for errors in terminology, writing, and acronyms. Semantic integrity checks are conducted to ensure that the meaning of the translated question matches the meaning of the original question. This process is carried out by certified translation experts. Figure 1 presents the flow of dataset translation process.

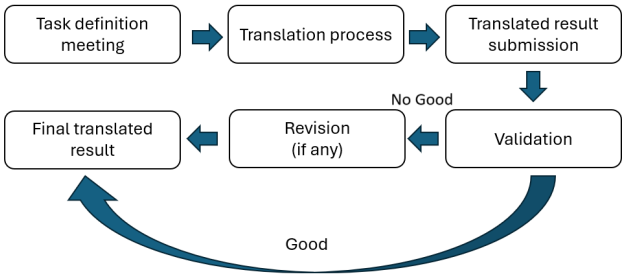

Figure 1. Flow of dataset translation process.

In the Task Definition Meeting stage, as illustrated in Figure 1, the preparation of the work tools is initiated. The required tools include the translation team, the original dataset files to be translated, storage locations, and tools for curation/annotation.

The translation team consists of a team leader and team members. The team leader is responsible for validating the translation results provided by the members. Meanwhile, the members are tasked with translating the original dataset into English and submitting it to the team leader. The team members are comprised of practitioners who translate from English to Indonesian, with two members holding bachelor's degrees in English literature and one member with a master's degree. The team leader position is held by a certified translator who is an expert in translating from English to Indonesian.

[3]https://github.com/AskNowQA/LC-QuAD2.0/tree/master/dataset

The team leader assigns translation tasks to the members. One team member translates the SimpleQuestions dataset, while three members translate LC-QuAD 2.0. Each question from the dataset is translated manually, one by one, with careful consideration of the corresponding sentences in Indonesian. The initial translation results are then submitted to the team leader. The team leader verifies and validates the first translation by considering several factors, including the accuracy of the translation, grammatical correctness, and appropriateness of the Indonesian equivalents used. Feedback from the verification and validation process is provided to the translation team members for revisions. Subsequently, the team members revise the initial translations based on the notes from the first verification and validation. Table 3 serves as a guideline for translators during the translation process.

Table 3. Translation Guideline

Check item	Caption
Entity	The entities mentioned in the questions and triples must be complete
Relation	The relations mentioned in the questions and triples must be complete
Grammar	The grammar used should be correct

An example of verification, as indicated by the three check items in Table 3, is the question taken from the SimpleQuestions test data: "Where did Roger Marquis die?". In this context, the triple for the answer is ("Roger Marquis" (Q7358590), "place of death" (P20), "Holyoke" (Q1637790)). Therefore, the entities and relations "Roger Marquis" and "*meninggal*" or "die" in English must be present in the translation. Additionally, the grammar must conform to correct Indonesian syntax. The accurate translation of the question should be "*Di mana Roger Marquis meninggal?*".

B. VALIDATION

In this section, the translation results are sent to the validator for validation. Table 4 serves as a guideline for the validator in conducting the validation of the translation results.

Table 4. Validation Guideline

Check item	Caption
Entity	The entities mentioned in the questions and triples must be complete
Relation	The relations mentioned in the questions and triples must be complete
Grammar	The grammar used should be correct
Meaning	The meaning of the translation must correspond to the original meaning

In the validation stage, the validator can utilize Table 4. This table is similar to Table 3 but includes an additional item, "Meaning", in its checks. During this stage, the validator conducts a re-examination of the "Entities" and "Relations" items that have been checked by the translators, as well as verifying the grammatical usage according to the validator's expertise. Likewise, for the "Meaning" item,

the validator is authorized to assess whether the translation accurately reflects the meaning of the original question. For example, a question from the SimpleQuestions test data, "What illness did Michael Visaroff die from?" was translated by the translator as "*Meninggal karena penyakit apa Michael Visaroff?*". The validator checks the "Entities", "Relations", and "Grammar" items as described in Table 3. However, for the "Meaning" item, the validator considers the translated meaning to be less accurate compared to the original question's meaning. Therefore, the validator modifies it to "*Penyakit apa yang menyebabkan Michael Visaroff meninggal?*".

In the subsequent stage, to ensure that the content of the dataset is minimally erroneous, we conduct a final checking process on all translated questions for both the SimpleQuestions and LC-QuAD 2.0 datasets. The final checking process involves data cleansing to correct writing errors and inaccuracies in the use of entity names. During this final checking process, we perform a manual review of each question within the SimpleQuestions and LC-QuAD 2.0 datasets, one by one. Table 5 presents the guidelines for the final checking stage.

Table 5. Final Checking Guideline

Check item	Caption
Writing	Writing must adhere to the General Guidelines for Indonesian Spelling (PUEBI).
Entity writing	The writing of entities must correspond to the entity labels in the triples

An example of the checks as outlined in Table 5 is taken from the SimpleQuestions dataset in the test data file. In the first check item, "Writing", a misspelling of the word "Menyutradari" is found, which should be corrected to "*Menyutradarai*" or "to direct" in English. For the second check item, "Entity Writing", the translator translated all words into Indonesian for the title of a film, specifically "the Tourist", which should have remained untranslated as "the Tourist" rather than being rendered as "Turis". When such errors occur, the translation is corrected to reflect the original entity name, which is "the Tourist".

V. RESULT AND DISCUSSION

In this section, the results and discussions related to the validation process of translations and final checking for each dataset, namely SimpleQuestions and LC-QuAD 2.0, are presented. The obtained results are divided into two sections. The first section summarizes the translation results of the SimpleQuestions data, while the second section addresses the results for LC-QuAD 2.0.

However, prior to this, the distribution of data for SimpleQuestions and LC-QuAD 2.0 can be observed in Table 6 [12]. Generally, the distribution of questions for each dataset is divided into training and testing data. However, for SimpleQuestions, in addition to the training and testing data, there is also validation data.

Table 6. The number of questions translated from SimpleQuestions and LC-QuAD 2.0

Dataset	Training data	Validation data	Testing data	Total
LC-QuAD2.0	22,132	-	9,689	31,821
SimpleQuestions	19,481	2,821	5,622	27,924

A. SIMPLEQUESTIONS

We conducted the translation process for 27,924 questions from SimpleQuestions. The stage following the translation by team members was the validation stage. During the validation stage, the validation was performed by certified translators. Based on the amount of data for each existing dataset, we implemented a sampling process for validation amounting to 10% [30]. The sampling was conducted randomly for each file within the datasets. This 10% sampling was performed to provide an overview of the quality of the translations prior to validation. The validation stage resulted in the tabulation presented in Table 7.

Table 7. Distribution of validated questions in SimpleQuestions.

Dataset	Training data	Validation data	Testing data	Total
SimpleQuestions	399	53	194	646

From Table 7 above, it can be observed that the initial translation results still require several corrections, totaling 646 questions. The majority of these corrections pertain to factoid-type questions that inquire about human entities but utilize the question word "what". For instance, the question "What baseball player was born in San Francisco?" was initially translated as "*Pemain baseball apa yang lahir di San Francisco?*" This translation was subsequently corrected to "*Siapa pemain baseball yang lahir di San Francisco?*".

Additionally, inaccuracies in the translation arose from the use of domain-specific verbs in the original questions. For example, the verb "directed" in the question "Who directed Doctor Dolittle? " was translated as "*Siapa yang mengarahkan Doctor Dolittle?*" However, the correct translation for "directed" should be "*yang menyutradarai*", thus the proper translation of the question should read "*Siapa yang menyutradarai Doctor Dolittle?*".

Other frequent errors occurred in questions requiring additional information that was not explicitly stated. For example, the question "Name a midfielder soccer player" includes the term "midfielder", which indicates a player's position in soccer. However, the question does not specify the word "position", leading to the translation "*Sebutkan pemain sepak bola gelandang?*". After correction, the translation becomes "*Sebutkan pemain sepak bola posisi gelandang?*" with the addition of the word "*posisi*" before "*gelandang*". In general, the accuracy of the translations prior to validation was 77%.

In the final checking stage, we identified several notes on SimpleQuestions. Table 8 presents the error statistics that

were found and corrected during the final checking stage.

Table 8. Error statistics identified and corrected during the final checking process for the SimpleQuestions dataset

Dataset	Training data	Validation data	Testing data	Total
Writing error	112	66	0	178
Entity Writing error	70	5	0	75

The checking results for two criteria from the SimpleQuestions dataset, as shown in Table 8, indicate a tendency for writing errors to occur more frequently than errors in entity naming, with occurrences of 178 and 75, respectively. This is logical given that the question type in the SimpleQuestions dataset comprises simple questions. These simple questions consist of a single triple, such as the question "*Dalam bahasa apa Mera Shikar difilmkan?*" which requires an answer derived from a single triple (Q6817891, P364, Q1568). In all cases of simple questions, only one named entity is utilized, namely "Mera Shikar". The entity comprises two words, {"Mera", "Shikar"}, while the non-entity words total four: {"*Dalam*", "*bahasa*", "*apa*", "*difilmkan*"}. Thus, the number of entity words is fewer than the number of non-entity words. Consequently, the tendency for translators to make errors in naming entities for these simple questions is lower.

B. LC-QUAD2.0

As shown in Table 6, the LC-QuAD 2.0 dataset consists solely of training and testing data, comprising 22,132 and 9,689 questions for the training and testing sets, respectively, resulting in a total of 31,821 questions.

For the validation process of LC-QuAD 2.0, we also employed a sample size of 10% from the total data available in LC-QuAD 2.0. The sampling was conducted randomly, and the validation process was carried out by certified translators. The 10% sampling in LC-QuAD 2.0 was intended to assess the quality of the translations prior to the validation process. Table 9 presents the results of the translation validation for LC-QuAD 2.0.

Table 9. Distribution of validated questions in LC-QuAD 2.0.

Dataset	Training data	Testing data	Total
LC-QuAD 2.0	1,392	212	1,604

In Table 9, the number of corrected questions is greater than that for SimpleQuestions, amounting to 1,604 questions. The translation errors identified during the validation process were largely similar to those found in the types of questions within SimpleQuestions. However, additional common errors encountered in the validation phase of LC-QuAD 2.0 pertained to grammar structure and meaning. This discrepancy arises from the dataset's inclusion of highly variable and complex question types. For instance, a question from the LC-QuAD 2.0 training data, "What is a sovereign state for the office held by the pope's head of state?", was corrected by the validator from "*Apakah negara berdaulat untuk jabatan yang dipegang oleh kepala negara paus?*" to "*Apa negara berdaulat untuk jabatan yang dipegang oleh kepala negara Paus?*". In both translations, the use of the interrogative word "Apakah" versus "Apa" carries distinct meanings. According to Table 9, overall, the translations deemed correct prior to validation exceeded 50%.

In the final checking process, several errors were identified based on the criteria outlined in Table 5. Table 10 presents the statistical findings from the final checking of the LC-QuAD 2.0 dataset.

Table 10. Error statistics identified and corrected during the final checking process for the LC-QuAD 2.0 dataset

Dataset	Training data	Testing data	Total
Writing error	88	188	196
Entity Writing error	178	6	184

The results of the checking on two criteria from the LC-QuAD 2.0 dataset, as illustrated in Table 10, indicate that, in general, the average errors observed in translation writing and entity naming are relatively similar, with counts of 196 and 184, respectively. This observation can be attributed to the predominance of complex questions within the LC-QuAD 2.0 dataset. Complex questions are defined as inquiries whose answers consist of a single triple. For instance, a question extracted from the LC-QuAD 2.0 training data, "*Kapan Prefektur Okinawa memiliki Departemen Santa Cruz sebagai badan administrasi kembarnya?*" necessitates three triples (wd:Q766445 p:P190 ?s . ?s ps:P190 wd:Q235106 . ?s pq:P580 ?value).

In this question, there are two entities: "Prefektur Okinawa" and "*Departemen* Santa Cruz". These two entities comprise five words, namely {"Prefektur", "Okinawa", "*Departemen*", "Santa", "Cruz"}. Meanwhile, the count of non-entity words totals six, namely {"*kapan*", "*memiliki*", "*sebagai*", "*badan*", "*administrasi*", "*kembarnya*"}.

A summary of the step-by-step construction process history of the Indo-WDSimpleQuAD2.0 dataset can be found in Figure 2.

Figure 2 presents a summary of the construction process journey for the Indo-WDSimpleQuAD2.0 dataset, from raw data (original dataset) to final data. Starting with the raw data, all questions were translated by translators, totaling 31,821 and 27,924 questions for LC-QuAD 2.0 and SimpleQuestions, respectively. At the validation stage, 10% of the total data was reviewed, resulting in 1,604 and 646 questions revised for LC-QuAD 2.0 and SimpleQuestions. Meanwhile, the final checking process for 100% of the data led to the correction of 380 and 253 questions for LC-QuAD 2.0 and SimpleQuestions, respectively.

To measure the consistency of annotations from different annotators, the inter-annotator agreement metric is used. This measurement samples 10% of the total questions,

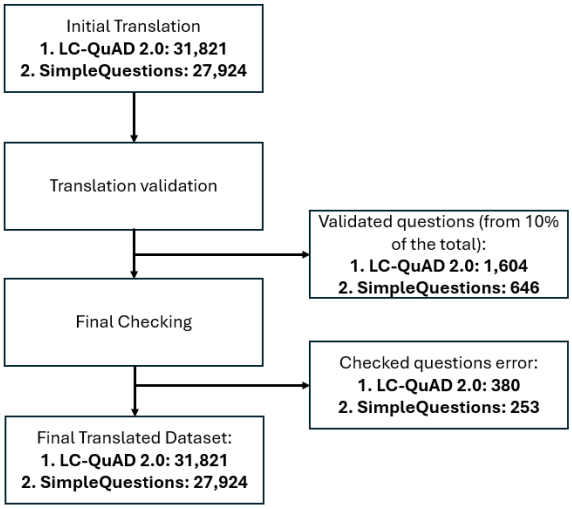

Figure 2. Summary of question validation and checking.

amounting to 2,791 questions for SimpleQuestions and 3,022 questions for LC-QuAD 2.0. We used Cohen's Kappa to measure inter-translator reliability [31].

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

P_o represents the accuracy, or the proportion of instances where the two raters gave identical labels. It is computed as:

$$P_o = \frac{(TP + TN)}{N} \tag{2}$$

P_e represents the likelihood that both raters would select the same label purely by chance. It is calculated as the sum of two components: P_1, the probability that both raters randomly select the first label (E) for Equivalent, and P_2, the probability that both choose the second label (I) for Inequivalent. These probabilities can be determined using the counts of true positives and true negatives previously mentioned, along with two additional terms:

$$P_1 = \frac{(TP + FN) * (TP + FP)}{N^2} \tag{3}$$

$$P_2 = \frac{(TN + FN) * (TN + FP)}{N^2} \tag{4}$$

From equations 1, 2, 3, and 4 above, the kappa values for the SimpleQuestions and LC-QuAD 2.0 datasets can be calculated. Table 11 and Table 12 are confusion matrices for 2,794 SimpleQuestions translations and 3,022 LC-QuAD 2.0 translations, respectively. From Table 11, we obtain the Kappa value for the SimpleQuestions dataset as follows:

$$\kappa = \frac{0.9312 - 0.7928}{1 - 0.7928} = 0.6681 \tag{5}$$

Table 11. Confusion Matric for the 2,791 Translation SimpleQuestions of Two Translators

	Translator 2 (E)	**Translator 2 (I)**	**Total**
Translator 1 (E)	2,368	58	2,426
Translator 1 (I)	134	231	365
Total	2,502	289	2,791

Cohen's Kappa value of approximately 0.6681 indicates substantial agreement between the two translators. Although the raw agreement appears strong, kappa reveals how much of that is beyond chance. The high value implies a consistent and reliable classification process between the two translators.

Table 12. Confusion Matric for the 3,022 Translation LC-QuAD 2.0

	Translator 2 (E)	**Translator 2 (I)**	**Total**
Translator 1 (E)	1,600	538	2,138
Translator 1 (I)	500	384	884
Total	2,100	922	3,022

Table 12 shows the confusion matrix for LC-QuAD 2.0 translations by two translators. From Table 12, the resulting Kappa score is as follows:

$$\kappa = \frac{0.6567 - 0.5780}{1 - 0.5780} = 0.1865 \tag{6}$$

A Cohen's Kappa value of approximately 0.1865 indicates slight to fair agreement between the two translators. While the observed agreement (66.57%) may seem high at first glance, this metric adjusts for the agreement that could occur by chance. Therefore, this result suggests that although there is some alignment, further refinement of criteria or training may improve inter-rater reliability.

VI. CONCLUSION

We developed a dataset called Indo-WDSimpleQuAD2.0, which consists of SimpleQuestions and LC-QuAD 2.0 in the Indonesian language, adhering to a silver standard. This dataset comprises two components: SimpleQuestions and LC-QuAD 2.0, containing 27,924 and 31,821 questions, respectively. Each dataset includes training, testing, and validation data, specifically for SimpleQuestions. The dataset development process involved the manual translation of the original English data into Indonesian by English language experts and native Indonesian speakers. The translated results were subsequently validated by certified translators (experts) to enhance the quality of the translations. Additionally, a final checking process was conducted to ensure that there were no errors in writing and naming entities in each question. We hope that the Indo-WDSimpleQuAD2.0

dataset will be valuable for researchers in the KGQA system who wish to conduct their studies using Indonesian-language data.

VII. ACKNOWLEDGEMENTS

This research is funded by Pusat Penelitian dan Pengabdian Kepada Masyarakat Politeknik Negeri Indramayu with the number of contract: 0666/PL42.PL42.9/AL.04/2025.

VIII. LIMITATIONS

In terms of translation scalability, the limitation of this research is the limited exploration of labels in the KG for each entity present in the question. In this case, the translator only translates the original question into Indonesian based on the words that form the question, without deeply considering whether the words refer to KG labels—specifically, whether the words are entities that should not be translated or not. However, by default, entity names such as people's names, places, countries, cities, foods, beverages, song titles, movie titles, and other known objects are left untranslated or not translated.

Regarding the applicability of the dataset, Indo-WDSimpleQuAD2.0 cannot be directly applied to other KGs but can still be used with slight adjustments to the SPARQL queries to obtain answers. This is because, although the facts in other KGs are the same as those in Wikidata, the forming triples may differ. For example, the fact about Joko Widodo, the 7th President of Indonesia and a politician from the Indonesian Democratic Party of Struggle (PDIP), is the same in both Wikidata and DBPedia. However, the properties used differ between Wikidata and DBPedia. Wikidata uses the property "member of political party," while DBPedia uses the property "party."

References

[1] F. Manola, E. Miller, and B. McBride, Eds., RDF 1.1 Primer. Cambridge, MA, USA: W3C Recommendation, 24 June 2014. [Online]. Available: https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

[2] M. I. Rahajeng and A. Purwarianti, "Indonesian question answering system for factoid questions using face beauty products knowledge graph," Jurnal Linguistik Komputasional, vol. 4, no. 2, pp. 59–63, September 2021. [Online]. Available: https://inacl.id/journal/index.php/jlk/article/view/62/46

[3] D. Kerenza and A. A. Krisnadhi, "Ac-iquad: Automatically constructed indonesian question answering dataset by leveraging wikidata," Lang Resources & Evaluation, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10579-023-09702-y

[4] L. Zhang, J. Zhang, X. Ke, H. Li, X. Huang, Z. Shao, S. Cao, and X. Lv, "A survey on complex factual question answering," AI Open, vol. 4, pp. 1–12, 2023. [Online]. Available: https://doi.org/10.1016/j.aiopen.2022.12.003

[5] M. Yani, A. A. Krisnadhi, and I. Budi, "A better entity detection of question for knowledge graph question answering through extracting position-based patterns," J. Big Data, vol. 9, no. 1, p. 80, 2022. [Online]. Available: https://doi.org/10.1186/s40537-022-00631-1

[6] P. J. Ochieng, "PAROT: translating natural language to SPARQL," Expert Syst. Appl., vol. 176, p. 114712, 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.114712

[7] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A. N. Ngomo, "Survey on challenges of question answering in the semantic web," Semantic Web, vol. 8, no. 6, pp. 895–920, 2017. [Online]. Available: https://doi.org/10.3233/SW-160247

[8] M. Yani and A. A. Krisnadhi, "Challenges, techniques, and trends of simple knowledge graph question answering: A survey," Inf., vol. 12, no. 7, p. 271, 2021. [Online]. Available: https://doi.org/10.3390/info12070271

[9] S. Pramanik, J. Alabi, R. S. Roy, and G. Weikum, "UNIQORN: unified question answering over RDF knowledge graphs and natural language text," CoRR, vol. abs/2108.08614, 2021. [Online]. Available: https://arxiv.org/abs/2108.08614

[10] N. Steinmetz and K. Sattler, "What is in the KGQA benchmark datasets? survey on challenges in datasets for question answering on knowledge graphs," J. Data Semant., vol. 10, no. 3-4, pp. 241–265, 2021. [Online]. Available: https://doi.org/10.1007/s13740-021-00128-9

[11] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," CoRR, vol. abs/1506.02075, 2015. [Online]. Available: http://arxiv.org/abs/1506.02075

[12] A. A. Krisnadhi, M. Yani, and I. Budi, "Entity and relation linking for knowledge graph question answering using gradual searching," Jurnal Nasional Teknik Elektro dan Teknologi Informasi, vol. 13, no. 2, pp. 139–146, 2024. [Online]. Available: https://jurnal.ugm.ac.id/v3/JNTETI/article/view/9184

[13] M. Azmy, P. Shi, J. Lin, and I. F. Ilyas, "Farewell freebase: Migrating the simplequestions dataset to dbpedia," in Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Association for Computational Linguistics, 2018, pp. 2093–2103. [Online]. Available: https://aclanthology.org/C18-1178/

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[15] M. Dubey, D. Banerjee, A. Abdelkawi, and J. Lehmann, "Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia," in The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part II, ser. Lecture Notes in Computer Science, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. F. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon, Eds., vol. 11779. Springer, 2019, pp. 69–78. [Online]. Available: https://doi.org/10.1007/978-3-030-30796-7_5

[16] D. Lukovnikov, A. Fischer, and J. Lehmann, "Pretrained transformers for simple question answering over knowledge graphs," in The Semantic Web - ISWC 2019 - 18th International Semantic Web Conference, Auckland, New Zealand, October 26-30, 2019, Proceedings, Part I, ser. Lecture Notes in Computer Science, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. F. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon, Eds., vol. 11778. Springer, 2019, pp. 470–486. [Online]. Available: https://doi.org/10.1007/978-3-030-30793-6_27

[17] C. Unger, L. Bühmann, J. Lehmann, A. N. Ngomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012, A. Mille, F. Gandon, J. Misselis, M. Rabinovich, and S. Staab, Eds. ACM, 2012, pp. 639–648. [Online]. Available: https://doi.org/10.1145/2187836.2187923

[18] X. Huang, J. Zhang, D. Li, and P. Li, "Knowledge graph embedding based question answering," in Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019, J. S. Culpepper, A. Moffat, P. N. Bennett, and K. Lerman, Eds. ACM, 2019, pp. 105–113. [Online]. Available: https://doi.org/10.1145/3289600.3290956

[19] W. Zhao, T. Chung, A. K. Goyal, and A. Metallinou, "Simple question answering with subgraph ranking and joint-scoring," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 324–334. [Online]. Available: https://doi.org/10.18653/v1/n19-1029

[20] D. Luo, J. Su, and S. Yu, "A bert-based approach with relation-aware attention for knowledge base question answering," in 2020 International Joint Conference on Neural Networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020. IEEE, 2020, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN48605.2020.9207186

[21] E. Cabrio, P. Cimiano, V. López, A. N. Ngomo, C. Unger, and S. Walter, "QALD-3: multilingual question answering over linked data," in Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013, ser. CEUR Workshop Proceedings, P. Forner, R. Navigli, D. Tufis, and N. Ferro, Eds., vol. 1179. CEUR-WS.org, 2013. [Online]. Available: https://ceur-ws.org/Vol-1179/CLEF2013wn-QALD3-CabrioEt2013.pdf

[22] C. Unger, C. Forascu, V. López, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter, "Question answering over linked data (QALD-4)," in Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij, Eds., vol. 1180. CEUR-WS.org, 2014, pp. 1172–1180. [Online]. Available: https://ceur-ws.org/Vol-1180/CLEF2014wn-QA-UngerEt2014.pdf

[23] C. Unger, C. Forascu, V. López, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter, "Question answering over linked data (QALD-5)," in Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015, ser. CEUR Workshop Proceedings, L. Cappellato, N. Ferro, G. J. F. Jones, and E. SanJuan, Eds., vol. 1391. CEUR-WS.org, 2015. [Online]. Available: https://ceur-ws.org/Vol-1391/173-CR.pdf

[24] A. Perevalov, D. Diefenbach, R. Usbeck, and A. Both, "Qald-9-plus: A multilingual dataset for question answering over dbpedia and wikidata translated by native speakers," in 16th IEEE International Conference on Semantic Computing, ICSC 2022, Laguna Hills, CA, USA, January 26-28, 2022. IEEE, 2022, pp. 229–234. [Online]. Available: https://doi.org/10.1109/ICSC52841.2022.00045

[25] R. Usbeck, A. N. Ngomo, B. Haarmann, A. Krithara, M. Röder, and G. Napolitano, "7th open challenge on question answering over linked data (QALD-7)," in Semantic Web Challenges - 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28 - June 1, 2017, Revised Selected Papers, ser. Communications in Computer and Information Science, M. Dragoni, M. Solanki, and E. Blomqvist, Eds., vol. 769. Springer, 2017, pp. 59–69. [Online]. Available: https://doi.org/10.1007/978-3-319-69146-6_6

[26] Kartik, F. Shenoy, D. Ilievski, D. Garijo, P. Schwabe, and Szekely, "A study of the quality of wikidata," Journal of Web Semantics, vol. 72, no. -, pp. 1–10, April 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1570826821000536

[27] L. Kaffee and E. Simperl, "Analysis of editors' languages in wikidata," in Proceedings of the 14th International Symposium on Open Collaboration, OpenSym 2018, Paris, France, August 22-24, 2018. ACM, 2018, pp. 21:1–21:5. [Online]. Available: https://doi.org/10.1145/3233391.3233965

[28] K. Tharani, "Much more than a mere technology: A systematic review of wikidata in libraries," The Journal of Academic Librarianship, vol. 47, pp. –, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0099133321000173

[29] S. Cahyawijaya, H. Lovenia, A. F. Aji, G. I. Winata, B. Wilie, F. Koto, R. Mahendra, C. Wibisono, A. Romadhony, K. Vincentio, J. Santoso, D. Moeljadi, C. Wirawan, F. Hudi, M. S. Wicaksono, I. H. Parmonangan, I. Alfina, I. F. Putra, S. Rahmadani, Y. Oenang, A. A. Septiandri, J. Jaya, K. D. Dhole, A. A. Suryani, R. A. Putri, D. Su, K. Stevens, M. N. Nityasya, M. F. Adilazuarda, R. Hadiwijaya, R. Diandaru, T. Yu, V. Ghifari, W. Dai, Y. Xu, D. Damapuspita, H. A. Wibowo, C. Tho, I. M. K. Karo, T. Fatyanosa, Z. Ji, G. Neubig, T. Baldwin, S. Ruder, P. Fung, H. Sujaini, S. Sakti, and A. Purwarianti, "Nusacrowd: Open source initiative for indonesian NLP resources," in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 13 745–13 818. [Online]. Available: https://doi.org/10.18653/v1/2023.findings-acl.868

[30] G. D. Israel, "Determining sample size," April 2009. [Online]. Available: https://www.psycholosphere.com/Determining%20sample%20size%20by%20Glen%20Israel.pdf

[31] B. D. Eugenio and M. Glass, "The kappa statistic: A second look," Comput. Linguistics, vol. 30, no. 1, pp. 95–101, 2004. [Online]. Available: https://doi.org/10.1162/089120104773633402

DR. MOHAMMAD YANI an assitant professor at Politeknik Negeri Indramayu, Indonesia. He is also as a Head for Research and Community Service at Politeknik Negri Indramayu. His interest research include knowledge graph question answering and natural language processing.

WAWAN SETIAWAN, S.PD., M.A a lecturer at Universitas Singaperbangsa Karawang, Indonesia. His interest research include materials development, TEFL, and second language acquisition.

RIZKY FRIHATMAWATI, M.PD a lecturer at Universitas Bina Sarana Informatika, Indonesia. Her interest research include english language learning.

WALI ATMAMUDIN a translator at Excel Translation (a sworn translator), Indonesia.

MUHAMAD MUSTAMIIN, S.PD., M.KOM a lecturer at Politeknik Negeri Indramayu, Indonesia. His interest research include information retrieval and information system.

RENDI, S.KOM., M.KOM a lecturer at Politeknik Negeri Indramayu, Indonesia. His interest research include data science.

ESTI MULYANI, S.KOM., M.KOM a lecturer at Politeknik Negeri Indramayu, Indonesia. Her interest research include software engineering.

FACHRUL PRALIENKA BANI MUHAMAD, S.ST., M.KOM a lecturer at Politeknik Negeri Indramayu, Indonesia. His interest research include software engineering and data mining.

ADILA ALFA KRISNADHI, PH.D a lecturer and researcher in the field of ontology at the Faculty of Computer Science, Universitas Indonesia. In addition to his roles as an educator and researcher, he is currently entrusted with the position of Research and Community Service Manager. His interest research include Ontology design patterns, semantic web, knowledge representation, automated reasoning, and machine learning.

PROF. DR. INDRA BUDI a full professor in computer science and information systems at the Faculty of Computer Science, Universitas Indonesia. His interest research include information extraction, text and data mining, and enterprise resource planning.