IndQNER: Named Entity Recognition Benchmark Dataset from the Indonesian Translation of the Quran

Ria Hari Gusmita[1,2](✉), Asep Fajar Firmansyah[1,2], Diego Moussallem[1,3], and Axel-Cyrille Ngonga Ngomo[1]

[1] Paderborn University, Warburger Street 100, Paderborn, Germany
firstname.lastname@uni-paderborn.de
https://dice-research.org/team/
[2] The State Islamic University Syarif Hidayatullah Jakarta, Ir. H. Juanda Street 95, Ciputat, South Tangerang, Banten, Indonesia
{ria.gusmita,asep.airlangga}@uinjkt.ac.id
[3] Jusbrasil, Brazil

Abstract. Indonesian is classified as underrepresented in the Natural Language Processing (NLP) field, despite being the tenth most spoken language in the world with 198 million speakers. The paucity of datasets is recognized as the main reason for the slow advancements in NLP research for underrepresented languages. Significant attempts were made in 2020 to address this drawback for Indonesian. The Indonesian Natural Language Understanding (IndoNLU) benchmark was introduced alongside IndoBERT pre-trained language model. The second benchmark, Indonesian Language Evaluation Montage (IndoLEM), was presented in the same year. These benchmarks support several tasks, including Named Entity Recognition (NER). However, all NER datasets are in the public domain and do not contain domain-specific datasets. To alleviate this drawback, we introduce IndQNER, a manually annotated NER benchmark dataset in the religious domain that adheres to a meticulously designed annotation guideline. Since Indonesia has the world's largest Muslim population, we build the dataset from the Indonesian translation of the Quran. The dataset includes 2475 named entities representing 18 different classes. To assess the annotation quality of IndQNER, we perform experiments with BiLSTM and CRF-based NER, as well as IndoBERT fine-tuning. The results reveal that the first model outperforms the second model achieving 0.98 F1 points. This outcome indicates that IndQNER may be an acceptable evaluation metric for Indonesian NER tasks in the aforementioned domain, widening the research's domain range.

Keywords: NER benchmark dataset · Indonesian · Spesific domain.

1 Introduction

Despite being the world's tenth most spoken language, with 198 million speakers[4], Indonesian remains an underrepresented language in the Natural Language Processing (NLP) field. The key issue for the slow progress in NLP for underrepresented languages is frequently defined as a lack of datasets [1]. Fortunately, major efforts to address the problem have recently been initiated: in 2020, the first Indonesian natural language understanding benchmark, IndoNLU, was created [9]. It includes benchmarks for 12 different core NLP tasks, which are divided into four categories: a) single-sentence classification, b) single-sentence sequence-tagging, c) sentence-pair classification, and d) sentence-pair sequence labeling. IndoBERT, an Indonesian pre-trained language model, was also introduced to enable the development of contextual language models in Indonesian.[5] The Indonesian Language Evaluation Montage (IndoLEM) was introduced in the same year as a comprehensive dataset for seven NLP tasks grouped into morpho-syntax and sequence labeling, semantics, and discourse coherence [4]. Another Indonesian pre-trained language model with the same name, IndoBERT, was also presented.[6] The most recent initiative (which is still ongoing at the time of writing) is to launch a joint-collaboration project named NusaCrowd.[7] The project's goal is to collect datasets written in Indonesian and its local languages and make them publicly available for reproducible research purposes.

Named Entity Recognition (NER) is one of the NLP tasks supported by the aforementioned benchmarks. NER is a fundamental task that identifies named entities (NEs) in unstructured or semi-structured text and assigns them to the appropriate classes. All the NER datasets were used to fine-tune IndoBERT. The results reveal that IndoBERT significantly increases the performance of the NER models. However, all datasets are intended for general use. This also applies to two versions of IndoBERT, as they were trained on formal and informal corpora from the general domain. In line with what has been done in [9] and [4], we propose IndQNER, an NER benchmark dataset for a specific domain, to help accelerate the advancement of NER research in Indonesian. Because Indonesia has the world's largest Muslim population[8], we choose the Indonesian translation of the Quran as a source for the dataset. This dataset is created using a meticulously designed annotation guideline, as well as the participation of Quran and Tafseer experts. It has 2475 NEs from 18 different classes. To properly measure the quality of the annotation, we conduct experiments in two different scenarios, including supervised and transfer learning. In the latter, we intend to discover how well IndoBERT can serve NER tasks in a specific domain. The evaluation results indicate that the dataset has the potential to significantly contribute to

[4] https://www.berlitz.com/blog/most-spoken-languages-world
[5] https://huggingface.co/indobenchmark/indobert-base-p1
[6] https://huggingface.co/indolem/indobert-base-uncased
[7] https://github.com/IndoNLP/nusa-crowd
[8] https://worldpopulationreview.com/country-rankings/muslim-population-by-country

broadening the domain range of Indonesian NER tasks and hence help to advance research development. IndQNER is now one of eight NER datasets that contribute to the NusaCrowd project.[9]

2 Related Work

Because there is no work on creating Indonesian NER datasets in a specific domain, we present the works in a general domain as follows.

The first attempt to create a NER dataset is described in [2]. This was motivated by the fact that the NER dataset generated by benefited Indonesian Wikipedia and DBpedia [6] contains numerous NEs with inaccurate labels. The main reason for this problem is the non-standard appearance of entities in DBpedia. Because the entity search applies exact matching, incomplete entities are disregarded and categorized as *Other*. To address this issue, a DBpedia entity expansion is proposed in order to produce a higher-quality dataset. All NEs are grouped into *Person*, *Location*, and *Organization* classes before performing name cleansing, normalization, expansion, and validation. The resultant NEs are evaluated using the Stanford NER library, and the obtained F1 score is more than twice that of [6].

Alfina et al. expanded the work in [2] to overcome incorrect assignments of NEs from the person class on Indonesian DBpedia. This issue contributed to the appearance of incorrect NEs and misplaced class members. Some rules from [2] are modified to correct *Person* entities to fix the problem. This produces a new set of entities known as MDEE (Modified DBpedia Entities Expansion). The changes to the rules include: 1) creating new rules for both removed and existing entries, and 2) revising existing rules. In addition, a new rule for the *Organization* class is added following a thorough analysis of its existing rules. Gazetteers are used to add country names and 100 city names to the MDEE after the Unicode-based names handling task obtains 500 new *Place* entities. All the rules and procedures result in over 6521 NEs for *Person*, and 2036 and 352 NEs for *Place* and *Organization*, respectively.

In 2020, [3] presented an Indonesian NER dataset that is claimed to be a more standardized Indonesian NER dataset.[10] The dataset is created by manually re-annotating an existing dataset, termed as S&N [8]. The annotation is performed by three native speakers. This is the first NER dataset created using an annotation guideline. The dataset is referred to as Idner-news-2k. The dataset consists of 2340 sentences and 5995 NEs from the *Person*, *Location*, and *Organization* classes.

3 The Indonesian Translation of the Quran

The Quran is the Muslim sacred book written in Arabic. It uses unique terms with distinct meanings. Furthermore, because the Quran is both rich in meaning

[9] `https://indonlp.github.io/nusa-catalogue/`
[10] `https://github.com/khairunnisaor/idner-news-2k`

and literary, translation into other languages becomes extremely difficult. We discuss the principles that were used to construct the Indonesian translation of the Quran in order to help readers understand the Quran easily. In particular, we examine the ideas from the NER viewpoint, i.e., how NEs, which are in the form of proper nouns, appear in the translation version. To acquire a clear understanding of the ideas, we contrast the Indonesian and English translations of the Quran.[11]

1. Some common nouns are provided with the corresponding proper nouns. In the English translation, this is accomplished by placing the article *the* before a common noun. In the Indonesian translation in Table 1, the proper noun *Sinai* was included in a bracket to indicate that the common noun *gunung* (mountain) refers to a mountain named Sinai.

Table 1: Some common nouns are presented along with their corresponding proper nouns in the Indonesian translation of the Quran.

Indonesian Translation	English Translation
Kami pun telah mengangkat gunung **(Sinai)** di atas mereka untuk (menguatkan) perjanjian mereka.182) Kami perintahkan kepada mereka, "Masukilah pintu gerbang (Baitulmaqdis) itu sambil bersujud". Kami perintahkan pula kepada mereka, "Janganlah melanggar (peraturan) pada hari Sabat." Kami telah mengambil dari mereka perjanjian yang kukuh.	And We raised over them **the mount** for [refusal of] their covenant; and We said to them, "Enter the gate bowing humbly", and We said to them, "Do not transgress on the sabbath", and We took from them a solemn covenant.

2. Non-popular proper nouns are paired with synonyms that are popular with readers. In Table 2, the proper noun *Ahmad* is a rare name, hence it is supplemented by its well-known synonym, *Nabi Muhammad* (Prophet Muhammad). In contrast, the English translation only mentions *Ahmad* and provides no further information.
3. Pronouns are supplemented by the proper nouns to which they refer. The proper nouns are written in brackets as additional information. The Indonesian translation in Table 3 contains a pronoun, i.e. *engkau* (you), followed by the proper noun it refers to, i.e. *Nabi Muhammad* (Prophet Muhammad). This is not the case in the English translation, as neither the proper noun nor the pronoun appear.

4 Methodology

4.1 Architecture and Workflow

IndQNER has a pipeline that includes 1) the definition of the initial classes and the corresponding NEs using the Quran concepts ontology (*Initials*), 2)

[11] All English translations are the sahih international version from `https://corpus.quran.com/translation.jsp`

Table 2: How non-popular proper nouns are presented in the Indonesian translation of the Quran.

Indonesian Translation	English Translation
(Ingatlah) ketika Isa putra Maryam berkata, "Wahai Bani Israil, sesungguhnya aku adalah utusan Allah kepadamu untuk membenarkan kitab (yang turun) sebelumku, yaitu Taurat, dan memberi kabar gembira tentang seorang utusan Allah yang akan datang setelahku yang namanya **Ahmad** (Nabi Muhammad)." Akan tetapi, ketika utusan itu datang kepada mereka dengan membawa bukti-bukti yang nyata, mereka berkata, "Ini adalah sihir yang nyata."	And [mention] when Jesus, the son of Mary, said, "O children of Israel, indeed I am the messenger of Allah to you confirming what came before me of the Torah and bringing good tidings of a messenger to come after me, whose name is **Ahmad**." But when he came to them with clear evidences, they said, "This is obvious magic."

Table 3: How some pronouns are presented in the Indonesian translation of the Quran.

Indonesian Translation	English Translation
Kebenaran itu dari Tuhanmu. Maka, janganlah sekali-kali **engkau** (Nabi Muhammad) termasuk orang-orang yang ragu.	The truth is from your Lord, so never be among the doubters.

the definition of the comprehensive annotation guideline, 3) text annotation, 4) the modification of the annotation guideline during text annotation, 5) the verification of the class and NE candidates by involving experts, and finally 6) the annotation of new NEs before producing the final annotated dataset, as can be seen in Figure 1.

4.2 Named Entities and Classes

Because the Quran is a holy book, we believed that the NEs and classes derived from it needed to be meticulously defined. Conveniently, there is a publicly available Arabic Quranic corpus with three levels of analysis required in computational linguistics tasks: morphological annotation, syntactic treebank, and semantic ontology.[12] The latter comprises Quranic main concepts, as well as instances of the lowest-level concepts.[13] Each instance is supplemented with a Quran verse in which it appears, so the reader can acquire a solid understanding of the instance in the Quran. We started defining NEs and classes for our dataset by examining the lowest-level concepts as well as the instances. If an instance is a proper noun, we set the corresponding name in the Indonesian translation of the Quran as an NE and the concept as the NE class. This resulted in the creation of initial NEs and classes for our dataset. We term them *Initials*, and the classes include *Allah*, *Artifact*, *Astronomical body*, *Event*, *Holy book*, *Angel*,

[12] `https://corpus.quran.com/`

[13] `https://corpus.quran.com/concept.jsp`

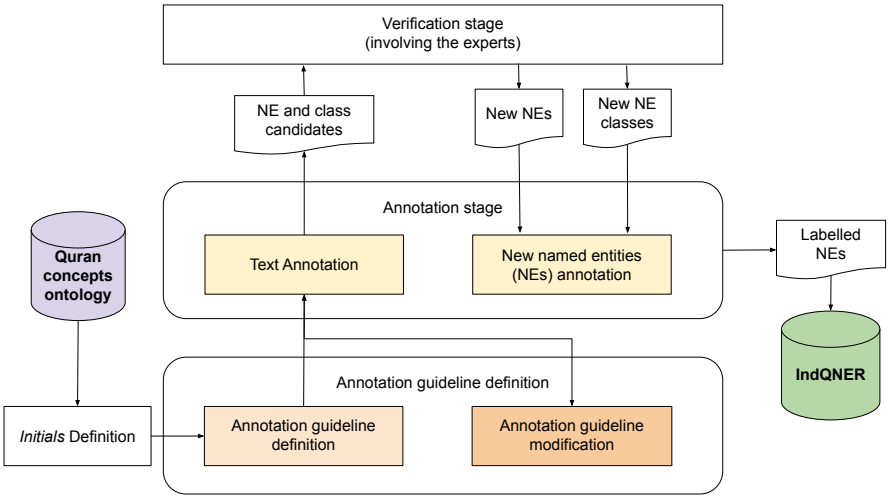

Fig. 1: The pipeline of IndQNER creation.

Person, *Location*, *Color*, and *Religion*. Following the trial annotation stage, we updated *Initials* to include more classes and NEs. The *Person* class is divided into three categories: *Messenger*, *Prophet*, and *Person*. The *Location* class is further classified into *Geographical location* and *Afterlife location*. We also added new classes, including *Allah's throne*, *False deity*, *Language*, and *Sentient*. We observed NEs that belong to two distinct classes: *Messenger* and *Prophet*. To address this, we transferred the NEs from the *Prophet* class to the *Messenger* class because while a messenger is certainly a prophet, not all prophets become messengers.

Since we discovered more NEs during the annotation stage, the number of NEs has increased (details are in Section 4.4). They are the synonyms of *Initials*' NEs, which typically appear in the following manners.

1. Synonyms appear as a name followed by an explanation in the form of an NE in *Initials*. The additional explanation is written in a pair of brackets. In Table 4, the Indonesian translation includes *Ruhulkudus* as a synonym of *Jibril* (an *Initials*' NE from the *Angel* class). The English translation just mentions *The Pure Spirit* instead of a name like *Jibril*.

Table 4: A synonym exists as a name followed by the corresponding NE in Initials that is written in a pair of brackets.

Indonesian Translation	English Translation
Katakanlah (Nabi Muhammad), "**Ruhulkudus** (Jibril) menurunkannya (Al-Qur'an) dari Tuhanmu dengan hak untuk meneguhkan (hati) orang-orang yang telah beriman dan menjadi petunjuk serta kabar gembira bagi orang-orang muslim (yang berserah diri kepada Allah)."	Say, [O Muhammad], "**The Pure Spirit** has brought it down from your Lord in truth to make firm those who believe and as guidance and good tidings to the Muslims."

2. Synonyms exist as a name without an explanation of the corresponding NE in *Initials*. According to Table 5, the Indonesian translation has a name, *fajar* (dawn), which is a synonym of an NE in *Initials*, *subuh* (dawn). The English version, by contrast, portrays *fajar* in a descriptive manner, i.e., *the white thread of morning distinguishes itself from the dark thread [of night].*

Table 5: A synonym exists as a name without additional information in the Indonesian Translation of the Quran.

Indonesian Version	English Version
Dihalalkan bagimu pada malam puasa bercampur dengan istrimu. Mereka adalah pakaian bagimu dan kamu adalah pakaian bagi mereka. Allah mengetahui bahwa kamu tidak dapat menahan dirimu sendiri, tetapi Dia menerima tobatmu dan memaafkanmu. Maka, sekarang campurilah mereka dan carilah apa yang telah ditetapkan Allah bagimu. Makan dan minumlah hingga jelas bagimu (perbedaan) antara benang putih dan benang hitam, yaitu **fajar**. ...	It has been made permissible for you the night preceding fasting to go to your wives [for sexual relations]. They are clothing for you and you are clothing for them. Allah knows that you used to deceive yourselves, so He accepted your repentance and forgave you. So now, have relations with them and seek that which Allah has decreed for you. And eat and drink until **the white thread of dawn becomes distinct to you from the black thread [of night]**. ...

3. The synonyms of *Allah* are precisely defined. They must be among the 99 names for Allah known as Asmaul Husna.[14] *Yang Maha Pengasih* (The Most or Entirely Merciful) and *Yang Maha Penyayang* (The Bestower of Mercy) are examples of Asmaul Husna. In addition, we discovered names that possess characteristics of Allah's synonyms but do not appear in Asmaul Husna. In this case, we used the Arabic version of the names and verified that they appeared in the Asmaul Husna reference before deciding they were a valid synonym of Allah. For example, in the Asmaul Husna reference, *Maha Mengurus* appears as *Maha Pemelihara* (The Guardian, The Witness, The Overseer). Based on our analysis of all appearances of Allah's synonyms in the translation of the Quran, we defined three forms of their existence in the text as follows:
   (a) One of Asmaul Husna's names that is preceded with word *Yang* (The). The appearance of *Yang Maha Pengasih* in this translation "*Sesungguhnya bagi orang-orang yang beriman dan beramal saleh, (Allah) Yang Maha Pengasih akan menanamkan rasa cinta (dalam hati) mereka.* (Indeed, those who have believed and done righteous deeds - the Most Merciful will appoint for them affection.)" is defined as a synonym of Allah.
   (b) Two names of Asmaul Husa that are preceded with word *Yang* and connected with word *lagi* (also). For example, *Yang Mahahalus lagi Mahateliti* in a translation "*Dia tidak dapat dijangkau oleh penglihatan mata, sedangkan Dia dapat menjangkau segala penglihatan itu. Dialah*

[14] We used the Asmaul Husna reference that can be seen at `https://github.com/dice-group/IndQNER/blob/main/Asmaul_Husna_Reference.pdf`

Yang Mahahalus lagi Mahateliti. (Vision perceives Him not, but He perceives [all] vision; and He is the Subtle, the Acquainted.)" is a synonym of Allah.

(c) One or two names of Asmaul Husna that is/are preceded with the phrase *Tuhan Yang* and connected with the word *lagi* (when two names exist). A phrase *Tuhan Yang Maha Penyayang* in "*(Ingatlah) Ayyub ketika dia berdoa kepada Tuhannya, "(Ya Tuhanku,) sesungguhnya aku telah ditimpa penyakit, padahal Engkau Tuhan Yang Maha Penyayang dari semua yang penyayang."* (And [mention] Job, when he called to his Lord, "Indeed, adversity has touched me, and you are the Most Merciful of the merciful.")" is defined as a synonym of Allah.

The annotation stage also produces candidates of NE and class. We consulted Quran and Tafseer experts to see if a candidate should be classified as an NE or a class (details are in Section 4.5).

4.3 Annotation Guideline

We designed the annotation guideline for IndQNER creation because there none existed for the domain. We had the preliminary version before the annotation. This version was updated during the annotation process based on findings in the Indonesian translation of the Quran (we refer to it as corpus) discovered during the annotation process. The guideline includes detailed instructions on how to annotate, what to annotate, and what information to collect during the annotation process. Each one is detailed in depth below.

How to do the annotation. The annotation is performed using Tagtog, a web-based text annotation tool.[15] Each of the two annotators labels two different chapters of the Indonesian translation of the Quran. Labeling is conducted by first selecting an NE and then specifying the appropriate label, as shown in Figure 2.

What to annotate. In the beginning, we have a list of NEs as well as the corresponding classes, as mentioned in Section 4.2. The annotators must locate NEs in the corpus and assign appropriate labels to them. *Person*, *Messenger*, and *Prophet* NEs have an additional labeling rule that excludes the title of a name (if available). In this translation, for example, "*Kebenaran itu dari Tuhanmu. Maka, janganlah sekali-kali engkau (Nabi Muhammad) termasuk orang-orang yang ragu* (Because the truth comes from your Lord, never be among the doubters)", *Nabi Muhammad* is an NE from *Messenger*. Because it appears with a title, *Nabi* (Prophet), the annotators should merely label *Muhammad*. Since synonyms of NEs are regarded as NEs, annotators must ascertain if a name that does not appear in *Initials* is a synonym of an NE. This is done by acquiring more information about the name from Wikipedia, either in Indonesian[16] or in English.[17] To validate Allah's synonyms, annotators must first verify if a name with relevant

[15] `https://www.tagtog.com/`

[16] `https://id.wikipedia.org/wiki/Halaman_Utama`

[17] `https://en.wikipedia.org/wiki/Main_Page`

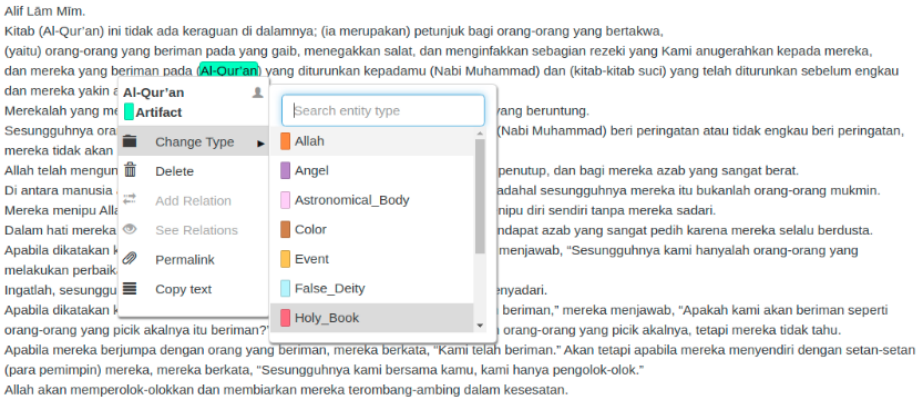

Fig. 2: The annotation process on Tagtog.

criteria exists in the Asmaul Husna reference. If no acceptable name is found, they need to find the Arabic version of the name and then check to ensure that the Arabic name is in the reference.

Which information to capture during the annotation process. According to the trial annotation stage's output, we discovered several names that did not exist in *Initials*. Therefore, the annotators are required to record these names and the location of their appearance in the corpus (including the chapter and verse numbers). Those names are considered NE candidates. In addition, the annotators might suggest a class candidate for each of the NE candidates. We also observed the presence of NEs in the form of metonymy in the corpus.[18] The annotators not only log these NEs, but also provide information about the classes involved. The annotators must also be aware if an NE in *Initials* belongs to the correct class. Furthermore, they must identify those that are improperly classified and recommend the correct one (if it is possible). All of this information is confirmed in the verification stage (Section 4.5).

4.4 Annotation Process and Results

The annotation process was carried out by eight annotators who are third- and fourth-year students at the Informatics Engineering Department of the State Islamic University Syarif Hidayatullah Jakarta. It was conducted in two stages, trial and actual, and held in two months. The trial step aimed to determine if all annotators have a common understanding of the annotation guideline. This stage also enabled us to discover several facts about the Indonesian translation of the Quran, as detailed in Sections 3 and 4.2. We used the Indonesian translation of the Quran that was released in 2019 by the Ministry of Religion Affairs of the

[18] `https://en.wikipedia.org/wiki/Metonymy`

Republic of Indonesia.[19] During the trial period, the annotators only annotated one chapter of the Quran, Al-Baqarah, with corresponding labels in *Initials*. We calculated the Inter-Annotator Agreement (IAA) among all annotators based on NE classes. *Color* class has the lowest number, with an IAA score of 3.7%. This is because two annotators labeled color names that do not appear in *Initials*. Those are actually intended to be NE candidates from the *Color* class. *Location* is the second class with an IAA score of less than 50%. This is because two annotators completely overlooked labeling location names.

To create IndQNER, we implemented the actual annotation step to annotate eight chapters in the Indonesian translation of the Quran. The Quran's chapters are classified as lengthy, medium, or short based on their number of words. We used seven lengthy chapters and one medium chapter. The lengthy chapters include *Chapter 2: Al-Baqarah* (l-baqarah), *Chapter 3: Ali-'Imran* (āl īm'rān), *Chapter 4: An-Nisā* (l-nisā), *Chapter 5: Al-Maidāh* (l-māidah), *Chapter 6: Al-An'ām* (l-an'ām), *Chapter 7: Al-A'rāf* (l-a'rāf), and *Chapter 10: Yūnus* (Yūnus). *Chapter 16: An-Nahl* (l-nahl) is a medium chapter. Each of the two annotators worked on two different chapters to conduct the annotation. Figure 3 shows the IAA scores obtained from annotation results across all chapters. NEs from *False Deity* and *Sentient* are nonexistent in all chapters. Meanwhile, the NEs from *Language*, *Afterlife Location*, and *Color* classes are each found in only one chapter, namely Chapter 16, Chapter 7, and Chapter 3, respectively. *Allah's Throne* and *Artifact* are two more classes whose NEs appear in less than half the number of chapters.

As mentioned in Section 4.3, annotators produce NE and class candidates in addition to the annotated corpus. We initially obtained 208 NE and three class candidates. After eliminating the duplicates, we had 142 NE and three class candidates left. We chose only NE candidates that are proper nouns to be checked in the next step. As a result, we had 59 NE and three class candidates. The majority of NE candidates were proposed as *Person*'s NEs. Furthermore, there are eight NE candidates whose proposed classes are still unknown. The names *Food* and *Fruit* were proposed for the two class candidates, but one class candidate's name remained unknown. The annotators also discovered one NE that was incorrectly classified, namely *Daud*. The annotators suggested *Messenger* as the correct class rather than keeping it as an NE from the *Prophet* class. The annotation results also assisted in locating NE synonyms in the corpus. We discovered 38 synonyms for *Allah* and eight synonyms for other NEs. The first appears in two forms, including being preceded by the phrases *Tuhan Yang* and *Yang*, which appear nine and 29 times, respectively.

4.5 Expert Curation

All NE and class candidates obtained through the annotation process were validated by three Quran and Tafseer experts. The experts are lecturers in the

[19] `https://lajnah.kemenag.go.id/unduhan/category/3-terjemah-al-qur-an-tahun-2019`

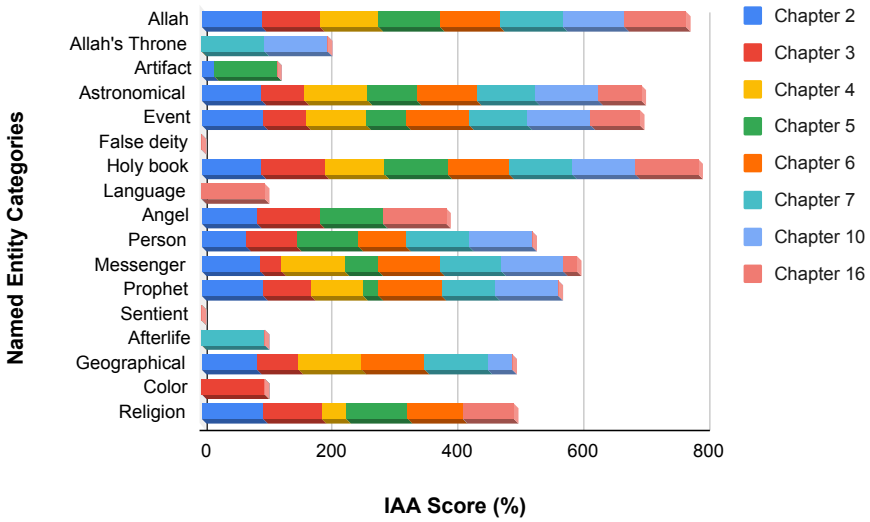

Fig. 3: Inter-annotator agreement (IAA) in the actual annotation stage.

Quran and Tafseer Department of the State Islamic University Syarif Hidayatullah Jakarta. To facilitate the verification process, we provided at least one Quran verse in which the NE candidates appear. Each expert specifically verified if the proposed NE classes were correct. In the case of unknown or incorrectly proposed NE classes, the expert provided the appropriate ones. Furthermore, the experts examined if the new proposed classes and corresponding NE candidates were acceptable. To obtain the final results, we chose the majority of the verification results. If each expert had a different result, we would ask them to discuss and decide on the final result. All experts shared the same results on 56 NE candidates. Meanwhile, two experts concluded the same results on three candidates. At this point, we had three more NE classes and 54 new NEs from existing classes. *Food*, *Fruit*, and *the Book of Allah* are the new classes, with two NEs for each of *Food* and *Fruit*, and one NE for *the Book of Allah*. Table 6 lists all classes, including the description and sample of the corresponding NEs.

5 Evaluation

Goals The purpose of the evaluation was to assess the annotation quality of IndQNER. In doing so, we acquired results from NER models testing, in which the models were trained on the dataset using two different settings. They were supervised learning and transfer learning, respectively. In the first setting, we used a combination of BiLSTM and CRF techniques [5] because it is the most commonly used approach in Indonesian NER tasks. An Indonesian pre-trained language model, IndoBERT,[20] was utilized to provide word embeddings. For the transfer learning setting, we used the IndoBERT which we fine-tuned and then tested. We were also interested in how much IndoBERT, which was trained on a

[20] `https://huggingface.co/indobenchmark/indobert-base-p1`

Table 6: All classes, descriptions, and samples of the corresponding named entities. The descriptions are taken from `https://corpus.quran.com/concept.jsp`.

Classes	Description	Sample of Named Entities
Allah	Allah (God in Islam) and all Allah's names that are known as Asmaul Husna	Allah, Tuhan Yang Maha Esa (*The Unique, The Only One*), Yang Maha Pengasih (*The Most or Entirely Merciful*)
Allah's Throne	The seat of Allah's power and authority	'Arasy (*Allah's Throne*)
Artifact	Man-made constructions that are mentioned in the Quran	Ka'bah (*Kaaba*), Masjidilaqsa (*Al-Aqsa mosque*)
Astronomical body	Astronomical objects that are mentioned in the Quran	Bintang Syi'ra (*Sirius*), bumi (earth)
Event	Temporal events	hari kiamat (*Day of Resurrection*), subuh (*fajr*)
False deity	The worhsip of false gods mentioned in the Quran	Al-'Uzza, Al-Lata (*al-Lat*)
Holy book	Holy books and other religious texts that are mentioned in the Quran	Al-Qur'an (*Qur'an*), Injil (*the Gospel*)
Language	The languages mentioned in the Quran	Bahasa Arab (*Arabic*)
Angel	The creations of Allah mentioned in the Quran known as angels	Malaikat maut (*The Angel of death*), Jibril (*Gabriel*)
Person	Individual human beings or groups of people mentioned in the Quran	Orang-orang Arab Badui (*The bedouins*), Azar (*Azar*)
Messenger	The messengers of Allah mentioned in the Quran	Ibrahim (*Abraham*), Muhammad (*Muhammad*)
Prophet	The prophets of Allah mentioned in the Quran	Harun (*Aaron*), Sulaiman (*Solomon*)
Sentient	The sentient creation mentioned in the Quran	makhluk bergerak dari bumi (*creature from the earth*)
Afterlife Location	Locations in the afterlife	Surga Firdaus (*The Gardens of Paradise*), Sidratulmuntaha (*The Lote Tree*)
Geographical location	Geographical locations mentioned in the Quran	Negeri Babilonia (*Babylon*), Makkah (*Makkah*)
Color	The different colors that are mentioned in the Quran	Hijau (*green*)
Religion	The major religions, or other systems of ancient belief, that are mentioned by name in the Quran	Islam (*Islam*), Nasrani (*Christianity*)
Food	The food mentioned in the Quran	Manna and Salwa
Fruit	The fruit mentioned in the Quran	Palm and Grave
The Book of Allah	The book of Allah mentioned in the Quran	Lauf Mahfuzh

large-scale general domain dataset, can support NER tasks in specific domains like the Indonesian translation of the Quran.

Experimental Setup The IndQNER dataset was annotated using the BIO (Beginning-Inside-Outside) tagging format. It has 3117 sentences, 62,027 tokens, and 2475 NEs. A sentence is marked by the end of a dot character. Each line in the dataset consists of a token, a tab, and the corresponding NE label. Figure 4 depicts the distribution of NEs in the dataset by class. *False Deity* and *Sentient* are two classes with no NEs in the corpus. To enable the two experiment settings, we split the dataset into training, validation, and test sets.[21] The split was made with an 8:1:1 ratio [7], with 2494, 312, and 311 sentences in the training, validation, and test sets, respectively.

Evaluations both in supervised and transfer learning settings were conducted with the following parameters: learning rate of 2e-5, maximum sequence length $\in \{256, 512\}$, batch size of 16, and number of epochs $\in \{10, 20, 40, 100\}$.

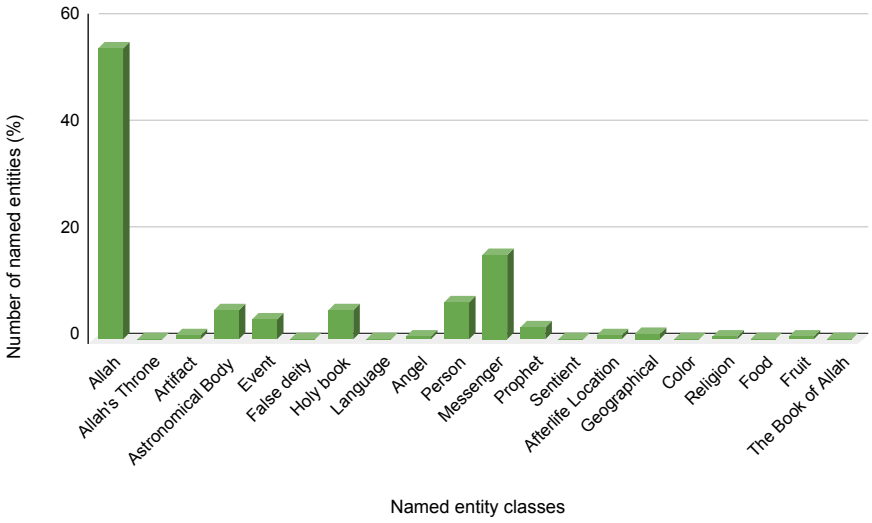

Fig. 4: Distribution of named entities from each class in IndQNER.

Results Table 7 provides evaluation results of IndQNER in two settings. The first NER model surprisingly outperforms the second on all setting parameters. The BiLSTM and CRF-based NER system obtains the highest F1 score of 0.98 and on other parameter settings, the numbers are consistently above 0.90. Meanwhile, the highest F1 score obtained from the fine-tuned IndoBERT model is 0.71. The results indicate that the existing Indonesian pre-trained language model is insufficient for supporting specific domain NER tasks, such as

[21] `https://github.com/dice-group/IndQNER/tree/main/datasets`

the Indonesian translation of the Quran. On the other hand, we believe that the annotation quality of the IndQNER dataset is satisfactory, as the learning process using a deep learning approach has been shown to successfully achieve a highly promising result.

Table 7: Evaluation results of IndQNER using supervised learning and transfer learning scenarios.

NER technique	e-poch 10			e-poch 20			e-poch 40			e-poch 100		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Max. sequence length 256												
Supervised learning	0.94	0.92	0.93	0.99	0.97	0.98	0.96	0.96	0.96	0.97	0.96	0.96
Transfer learning	0.67	0.65	0.65	0.60	0.59	0.59	0.75	0.72	0.71	0.73	0.68	0.68
Max. sequence length 512												
Supervised learning	0.92	0.92	0.92	0.96	0.95	0.96	0.97	0.95	0.96	0.97	0.95	0.96
Transfer learning	0.72	0.62	0.64	0.62	0.57	0.58	0.72	0.66	0.67	0.68	0.68	0.67

6 Conclusion and Future Works

We presented IndQNER, a NER benchmark dataset in a specific domain, namely the Indonesian translation of the Quran. This dataset creation is part of an attempt to satisfy the need for publicly accessible datasets in order to accelerate the progress of NLP research in Indonesian. The evaluation findings show that IndQNER can be a suitable metric for NER task evaluation in the Indonesian translation of the Quran domain. However, we are aware of the magnitude of IndQNER in comparison to the total number of chapters in the Quran. This is why we intend to grow the dataset to include all chapters in the future, so that there will be even more benefits available.

7 Acknowledgements

We acknowledge the support of the German Federal Ministry for Economic Affairs and Climate Action (BMWK) within the project SPEAKER (01MK20011U), the German Federal Ministry of Education and Research (BMBF) within the project KIAM (02L19C115) and the EuroStars project PORQUE (01QE2056C), and Mora Scholarship from the Ministry of Religious Affairs, Republic of Indonesia. Furthermore, we would like to thank our amazing annotators, including Anggita Maharani Gumay Putri, Muhammad Destamal Junas, Naufaldi Hafidhigbal, Nur Kholis Azzam Ubaidillah, Puspitasari, Septiany Nur Anggita, Wilda Nurjannah, and William Santoso. We also thank Khodijah Hulliyah, Lilik Ummi Kultsum, Jauhar Azizy, and Eva Nugraha for the valuable feedback.

References

1. Aji, A.F., Winata, G.I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R.E., Baldwin, T., Lau, J.H., Ruder, S.: One Country, 700+ Languages: {NLP} Challenges for Underrepresented Languages and Dialects in {I}ndonesia. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7226–7249. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.500, https://aclanthology.org/2022.acl-long.500
2. Alfina, I., Manurung, R., Fanany, M.I.: DBpedia entities expansion in automatically building dataset for Indonesian NER. 2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016 pp. 335–340 (2017). https://doi.org/10.1109/ICACSIS.2016.7872784
3. Khairunnisa, S.O., Imankulova, A., Komachi, M.: Towards a Standardized Dataset on {I}ndonesian Named Entity Recognition. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop. pp. 64–71. Association for Computational Linguistics (2020), https://www.tempo.co/https://www.aclweb.org/anthology/2020.aacl-srw.10
4. Koto, F., Rahimi, A., Lau, J.H., Baldwin, T.: {I}ndo{LEM} and {I}ndo{BERT}: A Benchmark Dataset and Pre-trained Language Model for {I}ndonesian {NLP}. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 757–770. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.66, https://www.aclweb.org/anthology/2020.coling-main.66
5. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1030, https://aclanthology.org/N16-1030
6. Luthfi, A., Distiawan, B., Manurung, R.: Building an Indonesian named entity recognizer using Wikipedia and DBPedia. Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014 pp. 19–22 (2014). https://doi.org/10.1109/IALP.2014.6973520
7. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information Extraction meets the Semantic Web: A Survey (2020). https://doi.org/10.3233/SW-180333, http://prefix.cc.
8. Syaifudin, Y., Nurwidyantoro, A.: Quotations identification from Indonesian online news using rule-based method. Proceeding - 2016 International Seminar on Intelligent Technology and Its Application, ISITIA 2016: Recent Trends in Intelligent Computational Technologies for Sustainable Energy pp. 187–194 (2017). https://doi.org/10.1109/ISITIA.2016.7828656
9. Wilie, B., Vincentio, K., Winata, G.I., Cahyawijaya, S., Li, X., Lim, Z.Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A.: IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding (9 2020), http://arxiv.org/abs/2009.05387