IndEL: Indonesian Entity Linking Benchmark Dataset for General and Specific Domains

Ria Hari Gusmita[1,2(✉)][0000−1111−2222−3333], Muhammad Faruq Amiral Abshar[2][0009−0002−7501−969X], Diego Moussallem[1,3][0000−0003−3757−2013], and Axel-Cyrille Ngonga Ngomo[1][0000−0001−7112−3516]

[1] Paderborn University, Warburger Street 100, Paderborn, Germany
ria.hari.gusmita@uni-paderborn.de,diego.moussallem@uni-paderborn.de, axel.ngonga@upb.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@uinjkt.ac.id,faruq.abshar19@mhs.uinjkt.ac.id

[3] Jusbrasil, Brazil

Abstract. In recent years, there has been a surge in natural language processing research focused on low-resource languages (LrLs), underscoring the growing recognition that LrLs deserve the same attention as high-resource languages (HrLs). This shift is crucial for ensuring linguistic diversity and inclusivity in the digital age. Despite Indonesian ranking as the 11th most spoken language globally, it remains underresourced in terms of computational tools and datasets. Within the semantic web domain, Entity Linking (EL) is pivotal, linking textual entity mentions to their corresponding entries in knowledge bases. This process is foundational for advanced information extraction tasks, including relation extraction and event detection. To bolster EL research in Indonesian, we introduce IndEL, the first benchmark dataset tailored for both general and specific domains. IndEL was manually curated using Wikidata, adhering to a rigorous set of annotation guidelines. We used two Named Entity Recognition (NER) benchmark datasets for entity extraction: NER UI for the general domain and IndQNER for the specific domain. IndQNER focused on entities from the Indonesian translation of the Quran. IndEL comprises 4765 entities in the general domain and 2453 in the specific domain. Using the GERBIL framework, we use IndEL to evaluate the performance of various EL systems, such as Babelfy, DBpedia Spotlight, MAG, OpenTapioca, and WAT. Our further investigation reveals that within Wikidata, a significant number of NIL entities remain unlinked due to the limited number of Indonesian labels and the use of acronyms. Especially in the specific domain, transliteration and translation processes performed to create the Indonesian translation of the Quran contribute to the presence of entities in a descriptive form and as synonyms.

Keywords: entity linking benchmark dataset · Indonesian · general and specific domains

1 Introduction

In recent years, there has been a surge in Natural Language Processing (NLP) research focused on low-resource languages (LrLs), underscoring the growing recognition that LrLs deserve the same attention as high-resource languages (HrLs). This shows that the shift is crucial for ensuring linguistic diversity and inclusivity in the digital age. In the rapidly evolving field of NLP, Entity Linking (EL) serves as a pivotal bridge, connecting raw textual mentions to structured entities within knowledge bases, such as DBpedia, YAGO, and Wikidata. Sentences *Affandi bergabung dengan kelompok Lima Bandung sekitar tahun 30-an* (Affandi joined Lima Bandung group around the 30s) and *Affandi berhasil memaksimalkan peran badan amil zakat kabupaten* (Affandi succeeded in bringing a district amil zakat foundation to its top performance) have one identical text mention i.e. *Affandi*. To help with the disambiguation of the two names, by leveraging Wikidata, EL can distinguish them. The first *Affandi* is recognized as a famous painter (wd:Q2826050), and the second one is identified as a regent (wd:Q20426359).

Indonesian, the lingua franca of the Indonesian archipelago, is spoken by over 278 million individuals[4]. Yet, the NLP resources tailored for it, remained underdeveloped until 2020. This is partly due to the absence of a robust benchmark dataset that can cater to both the general linguistic characteristics of Indonesian and its domain-specific nuances. Since 2020, noticeable efforts have been done to address the gap. Indonesian benchmark datasets for various NLP fundamental tasks in the general domain were presented along with the Indonesian pre-trained language model, IndoBERT [15,6]. More than 140 datasets for Indonesian NLP tasks were introduced as a result of the collaborative initiative to collect and unify existing resources for Indonesian languages [1]. Furthermore, [4] presents IndQNER as the first Named Entity Recognition (NER) benchmark dataset for Indonesian in a specific domain. However, to the best of our knowledge, no EL benchmark datasets are available for Indonesian both in general and specific domains.

To bridge this gap, we introduce IndEL, a meticulously crafted EL benchmark dataset tailored to Indonesian. We leveraged Wikidata, which encompasses a broad range of topics and domains, as the Knowledge Base (KB) to link entities within our dataset to their corresponding entries. IndEL caters to both general and specific domains, ensuring wide applicability across various use cases. Using NER UI[5], one of the Indonesian NER benchmark datasets, we extracted entities from the general domain, while IndQNER helped us identify entities in the specific domain[6]. The annotation process was conducted manually, adhering to a rigorous set of guidelines to ensure precision and consistency. IndEL contains 4765 and 2453 entities for general and specific domains, respectively. With the

[4] `https://www.worldometers.info/world-population/indonesia-population/#google_vignette`

[5] `https://github.com/indolem/indolem/tree/main/ner/data/nerui`

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

GERBIL benchmarking system, we use IndEL to evaluate cutting-edge EL systems, including Babelfy, DBpedia Spotlight, MAG, OpenTapioca, and WAT. Our evaluations underscore the dataset's potential as a foundational tool for advancing EL research in Indonesian, both in general and specific domains. Through this initiative, we contribute to advancing Indonesian as a developing-resource language.

2 Related Work

We outline some works in creating EL datasets in particular or multilingual settings.

[16] reported that the EL research for Chinese text is still in its early stages, and lacks publicly available annotated datasets and evaluation benchmarks. Existing Chinese corpora for EL are primarily constructed from noisy short texts, such as microblogs and news headings. Long texts, which represent a broader range of real-life scenarios, have been largely overlooked. The authors introduced CLEEK, a Chinese corpus of multi-domain long text for EL. CLEEK aims to promote the advancement of EL in languages other than English. CLEEK comprises 100 documents from various domains and is publicly accessible.

The first EL corpus for Icelandic was presented by [3]. Corpus annotation was conducted leveraging a multilingual entity linking model (mGENRE) combined with Wikipedia API Search (WAPIS). mGENRE is used to obtain record suggestions in Wikidata to expedite the EL labeling process in an Icelandic corpus. Meanwhile, WAPIS is leveraged to further enhance the labeling process, since it involves a search query run on the Wikipedia API. This method combination achieved a 53.9% coverage on the corpus, which was superior to the 30.9% coverage using only WAPIS.

In 2018, [12] presents the VoxEL dataset, a gold standard for EL in five European languages: German, English, Spanish, French, and Italian. The dataset is based on multilingual news, with 15 corresponding news articles for each language (75 articles in total). Two versions of VoxEL are created: a strict version focusing on traditional entity definitions (*Person*, *Place*, *Organization*) and a relaxed version considering a broader range of entities described by Wikipedia. Using the VoxEL dataset, the authors evaluate various EL systems to compare performance across systems and languages. They also compare the performance of EL systems for specific languages against results produced by translating the text to English using machine translation.

$\text{KORE}^{\text{DYWC}}$ was introduced as an extension of the KORE 50 data set to include YAGO, Wikidata, and Crunchbase [10]. The goal is to provide an evaluation data set that addresses the limitations of existing data sets and can be easily used by other developers. The KORE 50 data set was chosen as a foundation because it is popular and covers a broad range of topics in English. Three sub-data sets are released for each KB: YAGO, Wikidata, and Crunchbase. YAGO and Wikidata cover general knowledge, while Crunchbase focuses on technology and business. To perform the annotation, the authors used We-

bAnno, a web-based annotation tool, to manually annotate the KORE 50 data set using entities from different KBs. Each document was manually annotated by searching for entities in the respective KB. The annotations were exported using the WebAnno TSV3 format. There are some peculiarities of the annotation. Some entities were available in YAGO and Wikidata, but not in Crunchbase. YAGO offers a larger number of resources for annotation compared to DBpedia. Wikidata provides information for a broader range of mentions than DBpedia. Crunchbase has a tech-focused domain, resulting in fewer entities compared to DBpedia.

DocRED-FE was introduced as an English dataset that enhanced DocRED but with a redesigned entity type schema [14]. The new schema includes 11 coarse-grained types and 119 fine-grained types, providing richer contextual information. An example document is provided to illustrate the differences between the original DocRED and the new DocRED-FE schema. The authors used WebAnno for manual annotation, linking each entity to Wikidata to determine its types. The annotations were based on a new schema that was designed through a bottom-up, data-driven approach. The schema was refined through iterative exploratory annotation, with feedback from annotators leading to adjustments in the schema. Some entities were available in multiple types, and the authors had to make decisions on which type to assign based on context. The new schema was more precise and expressive compared to the original DocRED schema. The authors provide a comparison of DocRED-FE with other well-known datasets, highlighting the unique features of their dataset. They also analyze the distribution of entity types in their dataset, noting the top and least frequent types. The authors conducted experiments to evaluate JERE models on both DocRED and DocRED-FE. They found that DocRED-FE posed a greater challenge to existing models, but the fine-grained entity information improved relation classification performance.

3 Datasets Construction

In this section, we detail the development process of IndEL. We begin by discussing the document sources, from which we extracted entities for both the general domain (NER UI) and the specific domain (IndQNER). Subsequently, we shed light on the challenges posed by the entities from NER UI. We then delve into the crafting of the annotation guidelines, the manual annotation process, and the resultant findings.

3.1 Document Source of IndEL

Given the limited resources, we utilized two benchmark datasets for Indonesian NER, NER UI[7] and IndQNER[8], to obtain entities. NER UI and IndQNER are

[7] `https://github.com/indolem/indolem/tree/main/ner/data/nerui`
[8] `https://github.com/dice-group/IndQNER/tree/main/datasets`

designed to aid the benchmarking of Indonesian NER systems in general and specific domains, respectively. NER UI is from the news domain, and contains 5055 entities from *Person* (1870 entities), *Organization* (1949 entities), and *Location* (1236 entities) classes. Out of two Indonesian NER benchmark datasets introduced in 2020, NER UI has been shown as the best dataset as IndoBERT fine-tuning performed with it yields the highest F1 score of 90.1% [6]. Meanwhile, IndQNER is the first Indonesian NER benchmark dataset in a specific domain, the Indonesian translation of the Quran. It was presented with 3117 sentences and 2475 entities from 18 entity classes as explained in [4]. An evaluation of BiLSTM and CRF-based Indonesian NER system performed with IndQNER obtains an F1 score of 98% [4].

3.2 Challenges from Document Source

The NER UI dataset, while valuable as a document source for the general domain, presents several challenges that can impact the performance of EL systems. These challenges can be categorized into misspelled entities, incorrect entity spans, and missing entities.

Misspelled entities - Misspellings in entity names can hinder the ability of EL systems to correctly identify and link them to the appropriate entries in KBs. Table 1 provides examples of such misspellings from the NER UI dataset. The entities *Lea Iacocca* (the first example) and *Lentang* (the second example) are incorrectly spelled and should be written *Lee Iacocca* and *lenteng*, respectively.

Incorrect entity spans - The dataset sometimes incorrectly labels spans of text as entities or fails to capture the full span of an entity. Table 2 showcases this issue. In the first example, *Fakultas Ekonomi* (Economics Faculty) is labeled as a common noun, while *Universitas Indonesia* (the University of Indonesia) is identified as a proper noun. However, in the given context, both entities should be combined to form a single entity: *Fakultas Ekonomi Universitas Indonesia* (Economics Faculty at the University of Indonesia). A similar issue arises with the entities *Pemkot* (city/local government) and *Surabaya* in the second example, which should be combined as *Pemkot Surabaya*.

Missing entities - There are instances where valid entities are entirely overlooked in the dataset. Table 3 highlights such omissions, including entities like *Hye-kyo* (*Person*) and *Korea Times* (*Organization*) in the first example, and *Kabinet Kerja* (*Organization*) in the second example.

3.3 Annotation Guidelines

To help the annotators with the same knowledge of how to do the annotation, we designed the annotation guidelines meticulously.[9] The guidance presents information pertaining to two aspects as follows.

[9] `https://github.com/dice-group/IndEL/blob/main/Annotation_guidelines_%20in_%20English.pdf`

Table 1: Examples of misspelled entities in the document source of the general domain dataset.

First Example	Second Example
<entity>**Lea Iacocca**</entity> mampu secara cepat membenahi <entity>Chrysler</entity> karena dia mempunyai wewenang penuh melakukan konsolidasi, termasuk membawa beberapa kolega lamanya dari <entity>Ford</entity>.	Itu bukan etika <entity>PDIP</entity>, ujar <entity>Hasto</entity> di sela pelatihan manajer kampanye kader <entity>PDIP</entity> di kantor DPP, <entity>Jl **Lentang** Agung</entity>, <entity>Jakarta Selatan</entity>, Kamis (7/4/2016).

Table 2: Examples of incorrect entity spans in the document source of the general domain dataset.

First Example	Second Example
Mantan Dekan <entity>**Fakultas Ekonomi**</entity> <entity>**Universitas Indonesia**</entity> ini mengatakan ...	... capaian yang sudah dilakukan <entity>Risma</entity> dan <entity>**Pemkot**</entity> <entity>**Surabaya**</entity> terhadap kepedulian ...

Table 3: Examples of missing entities in the document source of the general domain dataset.

First Example	Second Example
"Dia memerankan karakternya dengan sangat bagus , menarik, bahkan membuat saya berdebar," kata **Hye-kyo** lagi, yang dikutip oleh **Korea Times**, Rabu (20/4/2016).	Sinyal akan dilakukannya reshuffle **Kabinet Kerja** oleh Presiden <entity>Joko Widodo</entity>terus berhembus.

How to annotate The manual annotation is performed using a semantic annotation platform, INCEpTION [5]. Annotators are tasked with identifying entities within the text and associating them with the corresponding Wikidata entries. INCEpTION facilitates this process by allowing annotators to search for entities directly on Wikidata. It is crucial for annotators to verify that the links they find correspond accurately to the entities mentioned in the text and that these links include Indonesian labels.

What to annotate Before beginning the annotation process, all annotators are provided with two types of documents: one containing raw text with sentences and another with the same text pre-tagged with entities. The raw text serves as the workspace for annotators to locate and tag entities, while the pre-tagged document is intended to guide the annotators by highlighting the specific sections of text that are entities. Annotators can simply use the search function to link

entities to the correct entries on Wikidata if they are certain of the references based on the sentence context. If the names are incomplete or the context does not provide enough information for a confident identification, annotators are instructed to use Google's document retrieval function to search for the names within documents. If no relevant documents are found, the names remain untagged. To obtain correct links on Wikidata, annotators must disambiguate the entries by navigating them using the provided descriptions.

3.4 Human Annotation and Results

The manual annotation was initially carried out by six non-volunteer native speakers, with four annotators focusing on the general domain and two on another domain. Specifically, for the specific domain, the annotators were fourth-year bachelor's students from the Quran and Tafseer department at the State Islamic University Syarif Hidayatullah Jakarta. Each of the two annotators labeled the same document according to the designed annotation guidelines. Therefore, we had two groups of annotators labeling the general domain dataset, and one group was assigned to label the specific domain dataset. Furthermore, we had a third annotator that was tasked with verifying the annotation results manually. We started by conducting the trial annotation process to observe whether all annotators have the same understanding of the annotation process. In this stage, the annotators were asked to label all entities in 20 sentences from another Indonesian NER benchmark dataset, NER UGM. The actual annotation was done after all annotators demonstrated their common understanding of the annotation.

According to the analysis of the actual annotation results, we distinguished the labeled entities into three categories. They are *Agreed*, *Disagreed*, and *OneNoLink*. *Agreed* is used in the case when two annotators provide the same links for an NE. Different links from annotators will make an NE classified as *Disagreed*. When only one annotator provides a link for an NE, then it will be grouped in the last category, *OneNoLink*. This happens when another annotator does not think of the name as an NE, or overlooks it. Table 4 depicts the number of entities from all categories in both general and specific domains. In the case of the number of *OneNoLink* entities, we summed the number of entities that were annotated only by each of the annotators.

We performed the second annotation to resolve *OneNoLink* entities. We asked the respective annotators to relook at the document and decide whether the names must be labeled or remain as non-entities. At this point, although the number of entities in the *OneNoLink* group remained small, we obtained new entities in other groups. This altered the distribution of entities, as displayed in Table 5. To handle the remaining *OneNoLink* entities in the general domain, we first selected valid entities among them by checking whether the entities exist in the document source, NER UI. We obtained 114 and 122 valid entities from the first and second groups of annotators, respectively. The remaining entities were termed NE candidates. Both valid entities and NE candidates were presented with the link provided by the respective annotators in the second annotation

stage. The third annotator verified the NE candidates as well as the proposed links manually. In the case of valid entities, the third annotator just checked whether the proposed links were correct. Table 6 shows the results of the manual verification. In the section of *Valid Named Entities*, we term entities with correct proposed links as *Taken* where most annotators have the highest number of them. We also found valid entities that are actually common nouns, and thus we categorized them as *Invalid entities*. These entities mostly exist in the results of the first annotator in both groups. In the *Named Entity Candidates* section, *Taken* category is used to state NE candidates that were verified as valid entities and that the proposed links were correct. More than 59% of new entities could be identified by the majority of annotators. However, the first annotator in *Group 1* contributed the highest number of invalid entities. Furthermore, only the proposed links from the second annotator in *Group 1* needed to be corrected.

Table 4: Distribution of entities in Agreed, Disagreed, and OneNoLink categories for general and specific domains.

Domain	Agreed	Disagreed	OneNoLink
General-group 1	1975	246	527
General-group 2	1905	191	258
Specific	2266	179	34

Table 5: Distribution of entities in Agreed, Disagreed, and OneNoLink categories for general and specific domains after the second annotation.

Domain	Agreed	Disagreed	OneNoLink
General-group 1	2035	411	299
General-group 2	1905	276	191
Specific	2296	179	4

To resolve the *Disagreed* annotation results, the third annotator manually checked different proposed links on Wikidata from two annotators to determine the correct one. If no correct link was found, the annotator searched for the link manually, following the annotation guidelines. From this process, not only did we find the correct links, either from the proposed links or those suggested by the third annotator, but we also identified Not in Lexicon (NIL) and invalid entities. Table 7 describes the results of manual checking to handle *Disagreed* entities. Generally, we distinguished the checking results according to the source of the correct link. There are three categories of them, i.e. from one of the annotators, from the third annotator (term *New Link*), and no correct link available. The latter is divided into NIL entities and invalid ones. In *Group 1*, more than 50%

correct links were taken from the 2nd annotator, while the 1st annotator in *Group 2* contributed more than 64% of the correct links. In both groups, we had the same portion of entities with new links as many as 2.8%. Moreover, the existence of invalid entities in *Group 1* has much more portion than in *Group 2* where they numbered 18%.

Table 6: Manual verification results on valid entities and entity candidates for the general domain

Verification Results Category	Group 1		Group 2	
	1st Annotator	2nd Annotator	1st Annotator	2nd Annotator
Valid Named Entities				
Taken	77.9%	26.3%	73.1%	75%
NIL Entities	15.8%	57.9%	15.4%	6.25%
Invalid Entities	1%	36.8%	3.8%	15.26%
New Link	5.3%	15.8%	7.7%	3.1%
Named Entity Candidates				
Taken	70.6%	0.75%	75.9%	60%
Invalid Entities	25.5%	99.3%	24.1%	40%
New Link	3.9%	-	-	-

Table 7: Manual checking results on *Disagreed* category for general domain.

Results Checking Category	Group 1	Group 2
Taken from 1st annotator	25.4%	64.2%
Taken from 2nd annotator	52.5%	27.4%
New Link	2.8%	2.8%
NIL Entities	1.8%	4.2%
Invalid Entities	18%	1.4%

In the specific domain, the third annotator manually checked four entities that were still in the *OneNoLink* category and found no correct links for all of them. To handle entities in *Disagreed*'s, the annotator went through all two different proposed links and selected correct links as many as 73.2% from the 1st annotator and 26.8% from the 2nd annotator.

We applied the same procedure on entities in the *Agreed* category to maintain the annotation quality of IndEL. We first checked whether every NE was a valid one, and we found 119 entities do not appear in NER UI. Therefore,

we categorized them as NE candidates. The third annotator performed manual checking to determine whether the NE candidates were entities and whether the proposed links were correct. Finally, we collected 4765 and 2453 entities for general and specific domains, respectively. Details of the number of NIL entities and the number of entities affected by the challenges in the source document are provided in the repository. [10]

Deal with the challenges in the source dataset. To overcome challenges found in NER UI (Section 3.2), we extended the aim of manual verification that we have explained in Section 3.4. For example, when we performed a selection of two proposed links in *Disagreed* category, if we found no correct link we further checked if the entity falls in one issue in NER UI. If this is the case, we will make the appropriate corrections, such as finding the correct name (for misspelled entities), combining entities (for incorrect entities' span), and providing correct links on Wikidata (for missing entities).

Dataset analysis, format, and usage Table 8 presents the distribution of the number of unique entities, sentences with nested entities, and the average number of entities in each sentence in IndEL for general and specific domains. As expected, the general domain contains a substantially wider range of entities, as evidenced by the presence of 31% unique entities. It also has more sentences containing nested entities compared to the specific domain. An average appearance of 2.4 entities in sentences within the general domain denotes a more complex sentence structure than is typical in the specific domain. The lower number of unique entities and lower average number of entities per sentence in the specific domain support the fact that it has focused content. To meet the need for widely used EL benchmark datasets, IndEL was created in the NLP Interchange Format (NIF).[11] Furthermore, to facilitate the evaluation process of multilingual EL systems, IndEL has been integrated into the GERBIL platform [13]. This integration enables researchers to efficiently test and compare the performance of various EL systems across multiple languages.[12]

Table 8: Distribution of unique entities, sentence with nested entities, and entities in sentences.

Domain	Total Entities	Unique Entities	Sentence with Nested Entities	Entities in Sentence
General	4767	1488	55	2.4
Specific	2453	141	16	1.6

[10] https://github.com/dice-group/IndEL/tree/main
[11] https://github.com/dice-group/IndEL/tree/main/datasets
[12] https://gerbil.aksw.org/gerbil/

4 Experiments and Analysis

We performed experiments using IndEL to examine the performance of cutting-edge EL systems in multilingual contexts. These experiments aimed to understand how these EL systems operate and perform when dealing with the Indonesian language, thereby providing insights into their effectiveness and adaptability in diverse linguistic settings. In doing so, we use GERBIL, a framework that enhances the ease of comparing and analyzing different EL systems [13]. It allows a more uniform and efficient evaluation process by standardizing the way the systems are accessed, and their results are processed. GERBIL is also capable of translating identifiers across various KBs, ensuring compatibility and integration between different systems. We used micro-measures for precision, recall and F1 to show the performance over the set of all annotations inside the dataset.[13] From the systems integrated on GERBIL, only Babelfy [8], DBpedia Spotlight [7], MAG [9], OpenTapioca [2], and WAT [11] yielded results in our experiments.

Table 9 showcases the results of the experiments with all systems in both general and specific domains. It is observed that systems generally achieve greater precision within the specific domain compared to the general domain. DBpedia Spotlight excels in the specific domain but experiences a marked decline in its performance when used in the general domain. Conversely, OpenTapioca demonstrates superior performance in the general domain compared to the specific domain, where its precision outperforms all other systems. Babelfy maintains consistent precision across domains, but has a notable drop in recall when transitioning from the specific to the general domain. MAG shows a considerable increase in performance across all metrics when moving from the specific to the general domain. In contrast, WAT demonstrates considerably better results in the specific domain as compared to the general domain, particularly its F1 score, which surpasses all others. At this point, WAT emerges as the top-performing EL system for Indonesian text, securing the highest F1 score in both domains. These findings indicate that Indonesian entities in the Indonesian translation of the Quran may have more clear-cut entities, facilitating accurate identification by the systems. However, the inherent diversity of Indonesian entities in the general domain presents a greater challenge for multilingual EL systems. Furthermore, we provide details of the evaluation results as well as the performance of the mentioned EL systems on other benchmarks in the repository.[14]

To further investigate the impact of how Indonesian entities are presented on Wikidata on the performance of EL systems, we conducted an additional experiment using MAG, the EL system with the lowest F1 score both in general and specific domains. The experiment was aimed at the identification of NIL entities within both domains. Specifically, we explored if NIL entities are acknowledged as either entry names or as labels in the Indonesian language within Wikidata, which could potentially affect the EL systems' ability to correctly link entities.

[13] https://github.com/dice-group/gerbil/wiki/Precision,-Recall-and-F1-measure

[14] https://github.com/dice-group/IndEL/blob/main/README.md

Table 9: GERBIL evaluation of Babelfy, DBpedia Spotlight, MAG, OpenTapioca, and WAT in the general and specific domains of IndEL

Metrics	Babelfy	DBpedia Spotlight	MAG	OpenTapioca	WAT
General Domain					
Precision	0.7278	0.6750	0.4265	**0.7984**	0.6118
Recall	0.3719	0.3577	0.4166	0.4105	**0.5549**
F1	0.4923	0.4676	0.4215	0.5423	**0.5820**
Specific Domain					
Precision	0.8049	**0.8471**	0.1523	0.6179	0.7715
Recall	0.4725	0.6731	0.1508	0.0310	**0.7501**
F1	0.5954	0.7501	0.1515	0.0590	**0.7606**

For this purpose, we randomly selected 55 NIL entities from each domain. Our findings indicated that only 14.5% of NIL entities in the general domain are listed as entry names, with the number slightly lower at 10.9% in the specific domain. In contrast, 29.1% of NIL entities in the specific domain are represented as Indonesian labels, compared to only 12.7% in the general domain. The main reason for the scarce appearance of NIL entities as entry names and Indonesian labels on Wikidata in the general domain is the use of acronyms. In the specific domain, 22.2% of the 81.8% of NIL entities that do not appear as entry names exist as Indonesian labels. Additionally, approximately 60% of NIL entities that do not appear as entry names lack Indonesian labels because they are defined descriptively. For example, the entity *Allah* appears as *Tuhan dalam Islam* (God in Islam). Another reason is the use of corresponding synonyms for NIL entities in the Indonesian label section of Wikidata. Some examples are *Ummul Qura* vs. *Makkah* (Mecca), *Hari Akhir* vs. *Yaumul Qiyamah* (Qiyama), *Baitullah* vs. *Ka'bah* (Kaaba), and *Israil* vs. *Yaq̄ub* (Jacob in Islam). Based on the findings, several recommendations can be made to improve the performance of EL systems in linking Indonesian entities on Wikidata as follows:

1. Increasing the number of Indonesian labels for entities can enhance the accuracy of EL systems.
2. Standardizing terminology for entities, particularly those defined descriptively, will promote more consistent linking. This is especially relevant for entities in the specific domain due to the transliteration and translation process from the original Qur'an, which is written in Arabic, to the Indonesian translation.
3. Recognizing and incorporating synonyms in Indonesian will ensure comprehensive label inclusion.
4. Developing better methods to handle acronyms, especially in the general domain, will reduce the number of unlinked NIL entities.

5 Conlusion and Future Works

We have introduced a pioneering benchmark dataset specifically crafted to evaluate EL systems targeting the Indonesian language, covering both general and specific domains. The general domain entities were sourced from one of Indonesian NER benchmark datasets, NER UI. Meanwhile, IndQNER, which was built from the Indonesian translation of the Quran, was used to obtain entities in the specific domain. All entities in IndEL are provided with their corresponding links on Wikidata. A GERBIL benchmarking process demonstrates that IndEL can be employed as an appropriate evaluation metric for assessing the performance of EL systems in Indonesian, both in general and specific domains. However, we recognize the challenges posed by the limited scope of IndEL and the fact that many entities remain unlinked, primarily due to the insufficient quantity of Indonesian labels on Wikidata. To address the former, we plan to enrich the dataset by incorporating additional entities from various Indonesian NER benchmark datasets, such as NERGrit, NERP, NER UGM, etc.[15] To address the latter as well as to develop KB agnostic EL systems for Indonesian, we intend to broaden the range of entity links within IndEL to establish connections with other KBs, such as BabelNet, DBpedia and YAGO. This expansion aims to facilitate the seamless integration of Indonesian entities with broader semantic knowledge resources, contributing to improved accuracy and versatility of EL systems.

6 Acknowledgements

We acknowledge the support of the German Federal Ministry of Education and Research (BMBF) within the project COLIDE (01I521005D), the European Union's Horizon Europe research and innovation programme within the project ENEXA (101070305), the German Federal Ministry of Education and Research (BMBF) within the project KIAM (02L19C115), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the project SFB-TRR 318 (TRR 318/1 2021 – 438445824), and Mora Scholarship from the Ministry of Religious Affairs, Republic of Indonesia.

References

1. Cahyawijaya, S., Lovenia, H., Aji, A.F., Winata, G., Wilie, B., Koto, F., Mahendra, R., Wibisono, C., Romadhony, A., Vincentio, K., Santoso, J., Moeljadi, D., Wirawan, C., Hudi, F., Wicaksono, M.S., Parmonangan, I., Alfina, I., Putra, I.F., Rahmadani, S., Oenang, Y., Septiandri, A., Jaya, J., Dhole, K., Suryani, A., Putri, R.A., Su, D., Stevens, K., Nityasya, M.N., Adilazuarda, M., Hadiwijaya, R., Diandaru, R., Yu, T., Ghifari, V., Dai, W., Xu, Y., Damapuspita, D., Wibowo, H., Tho, C., Karo Karo, I., Fatyanosa, T., Ji, Z., Neubig, G., Baldwin, T., Ruder,

[15] `https://indonlp.github.io/nusa-catalogue/index.html`

S., Fung, P., Sujaini, H., Sakti, S., Purwarianti, A.: NusaCrowd: Open source initiative for Indonesian NLP resources. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 13745–13818. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.868, https://aclanthology.org/2023.findings-acl.868

2. Delpeuch, A.: Opentapioca: Lightweight entity linking for wikidata. ArXiv **abs/1904.09131** (2019), https://api.semanticscholar.org/CorpusID:125953443
3. Friðriksdóttir, S.R., Eggertsson, V.Á., Jóhannesson, B.G., Daníelsson, H., Loftsson, H., Einarsson, H.: Building an Icelandic entity linking corpus. In: Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference. pp. 27–35. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.dclrl-1.4
4. Gusmita, R.H., Firmansyah, A.F., Moussallem, D., Ngonga Ngomo, A.C.: IndQNER: Named Entity Recognition Benchmark Dataset from the Indonesian Translation of the Quran, vol. 2. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-35320-8{_}12, http://dx.doi.org/10.1007/978-3-031-35320-8_12
5. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9. Association for Computational Linguistics (June 2018), http://tubiblio.ulb.tu-darmstadt.de/106270/, event Title: The 27th International Conference on Computational Linguistics (COLING 2018)
6. Koto, F., Rahimi, A., Lau, J.H., Baldwin, T.: IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 757–770. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.66, https://aclanthology.org/2020.coling-main.66
7. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: Shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. p. 1–8. I-Semantics '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2063518.2063519, https://doi.org/10.1145/2063518.2063519
8. Moro, A., Raganato, A., Navigli, R.: Entity linking meets word sense disambiguation: a unified approach. Transactions of the Association for Computational Linguistics **2**, 231–244 (2014). https://doi.org/10.1162/tacl_a_00179, https://aclanthology.org/Q14-1019
9. Moussallem, D., Usbeck, R., Röeder, M., Ngomo, A.C.N.: Mag: A multilingual, knowledge-base agnostic and deterministic entity linking approach. In: Proceedings of the Knowledge Capture Conference. K-CAP 2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3148011.3148024, https://doi.org/10.1145/3148011.3148024
10. Noullet, K., Mix, R., Färber, M.: KORE 50ˆDYWC: An evaluation data set for entity linking based on DBpedia, YAGO, Wikidata, and crunchbase. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2389–2395. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.291

11. Piccinno, F., Ferragina, P.: From tagme to wat: A new entity annotator. In: Proceedings of the First International Workshop on Entity Recognition & Disambiguation. p. 55–62. ERD '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2633211.2634350, https://doi.org/10.1145/2633211.2634350
12. Rosales-Méndez, H., Hogan, A., Poblete, B.: Voxel: A benchmark dataset for multilingual entity linking. In: Vrandečić, D., Bontcheva, K., Suárez-Figueroa, M.C., Presutti, V., Celino, I., Sabou, M., Kaffee, L.A., Simperl, E. (eds.) The Semantic Web – ISWC 2018. pp. 170–186. Springer International Publishing, Cham (2018)
13. Verborgh, R., Röder, M., Usbeck, R., Ngonga Ngomo, A.C.: Gerbil – benchmarking named entity recognition and linking consistently. Semant. Web **9**(5), 605–625 (jan 2018). https://doi.org/10.3233/SW-170286, https://doi.org/10.3233/SW-170286
14. Wang, H., Xiong, W., Song, Y., Zhu, D., Xia, Y., Li, S.: Docred-fe: A document-level fine-grained entity and relation extraction dataset. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095786
15. 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. 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. pp. 843–857. Association for Computational Linguistics, Suzhou, China (Dec 2020), https://aclanthology.org/2020.aacl-main.85
16. Zeng, W., Zhao, X., Tang, J., Tan, Z., Huang, X.: CLEEK: A Chinese long-text corpus for entity linking. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2026–2035. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.249