MERLIN: A Testbed for Multilingual Multimodal Entity Recognition and Linking

Sathyanarayanan Ramamoorthy[◇] **Vishwa Shah**[◇] **Simran Khanuja**[◇] **Zaid Sheikh**[◇]
Shan Jie[†] **Ann Chia**[†] **Shearman Chua**[†] **Graham Neubig**[◇]
[◇]Carnegie Mellon University, USA
{sramamoo,vishwavs,skhanuja,zsheikh,gneubig}@cs.cmu.edu
[†]Defence Science and Technology Agency, Singapore
{yshanjie,achiajin,schuawei}@dsta.gov.sg

Abstract

This paper introduces MERLIN, a novel testbed system for the task of *Multilingual Multimodal Entity Linking*. The created dataset includes BBC news article titles, paired with corresponding images, in five languages: Hindi, Japanese, Indonesian, Vietnamese, and Tamil, featuring over 7,000 named entity mentions linked to 2,500 unique Wikidata entities. We also include several benchmarks using multilingual and multimodal entity linking methods exploring different language models like LLaMa-2 and Aya-23. Our findings indicate that incorporating visual data improves the accuracy of entity linking, especially for entities where the textual context is ambiguous or insufficient, and particularly for models that do not have strong multilingual abilities. For the work, the dataset, methods are available online.[1]

1 Introduction

Entity Linking (EL) involves linking ambiguous entity mentions from unstructured data to their unambiguous referents in a knowledge base (KB). This is an important step in many downstream applications such as search (Blanco et al., 2015), question-answering (Févry et al., 2020; De Cao et al., 2019), and sentiment analysis (Zhong et al., 2023). While traditional entity linking focused on text, in many settings (such as short news clips or social media posts), textual data alone can often be insufficient to adequately disambiguate entities given short, context-dependent, and ambiguous statements (Guo et al., 2013). Images, which often accompany text, can provide crucial context for disambiguating entities (Moon et al., 2018; Adjali et al., 2020). Multiple prior works have proposed datasets (Wang et al., 2022; Gan et al., 2021) and methods (Zhang et al., 2021) for EL given visual and textual context.

Additionally, EL presents unique challenges on non-English languages (Fu et al., 2020), where resources are much more constrained. In the multilingual context, earlier works framed this as a *cross-lingual* problem, where entities in one language were linked to a KB in another language, typically English (McNamee et al., 2011). Subsequent works (Botha et al., 2020; De Cao et al., 2021) utilized massively multilingual models with language-agnostic representations to link mentions to language-agnostic KBs in a multilingual setting.

Up until now, EL in the multilingual context has only dealt with the text modality. However, given the challenges of multilingual entity linking, it is natural to consider whether the visual modality may help resolve ambiguity in the multilingual case as well (as illustrated in Figure 1). The novelty of our work revolves around formulating the task of multilingual entity linking through multimodal data along with a dataset and baseline methods, the goal being to show the performance of existing EL methods on our dataset and also to encourage the community to work on disambiguating multilingual entity mentions using images. Therefore, to push the frontiers of EL towards this multilingual multimodal setting, we create **MERLIN**: the first testbed for **M**ultilingual Multimodal **E**ntity **R**ecognition and **LIN**king. MERLIN is a testset comprising news article titles from BBC, paired with corresponding images. It features over 7,000 named entity mentions linked to 2,500 unique entities in the Wikidata knowledge base (Table 2). The dataset covers five languages: Hindi, Japanese, Indonesian, Vietnamese, and Tamil.

In the following sections, we first formulate the task and describe the various components

[1]https://github.com/rsathya4802/merlin.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 399–417, 2026. https://doi.org/10.1162/TACL.a.633
Action Editor: Nitin Madnani. Submission batch: 9/2024; Revision batch: 4/2025; Published 4/2026.

Figure 1: *Resolving Entity Ambiguities Using Contextual and Visual Cues*. Entity disambiguation is challenging when multiple potential entities are associated with a single mention. Here, the ''Master'' films in different languages can each be linked to distinct IDs. Images serve as a vital tool allowing for correct linking of ''Master'' to its corresponding entity ID, thereby avoiding ambiguity.

involved in building a *multilingual multimodal entity linking* (**MMEL**) dataset in §2. Next, in §3, we describe the dataset selection and annotation process. In §4, we evaluate several methods that have been proposed for multilingual or multimodal entity-linking (De Cao et al., 2022; Shi et al., 2024), on our newly created test set. Analyzing results of these methods in §5, we find that even the best approaches lack in performance, demonstrating the difficulty of our task and dataset. We also show that *utilizing visual and text information* aids in disambiguating entities, leading to a significant improvement compared to methods that do not use visual information, especially for LLMs that aren't massively multilingual.

2 Task Formulation

Multilingual Multimodal Entity Linking (MMEL) is the task of mapping mentions m within multimodal and multilingual contexts to the corresponding entities in a KB. Formally, let $\mathcal{E}$ denote the entity set of the KB, typically comprising millions of entities e, each associated with a textual description T_e and an image V_e. Each mention m is characterized by its visual context V_m and textual context T_m, where V_m represents the image associated with m, and T_m represents the textual spans surrounding m and including m (e.g., article title and associated image).

Given an image V_m, corresponding textual context T_m, and a mention m embedded within T_m, the task of MMEL is to identify and link the mention m to its corresponding entity $e \in \mathcal{E}$, where $\mathcal{E}$ is defined as the set of all Wikidata entities in the KB. Formally, the objective is to output a set of mention-entity pairs: $\{(m_i, e_i)\}_{i=1}^{n_m}$, where each entity e_i corresponds to a mention m_i and n_m represents the total number of mentions in the given context.

In this multilingual setting, the context T_m can be presented in various languages, thereby necessitating robust cross-lingual and cross-modal understanding. The assumption is that each mention m_i has a valid corresponding entity e_i within the KB, the so-called in-KB evaluation paradigm. We do not consider the challenge of out-of-KB predictions in this work.

3 The Making of MERLIN

3.1 Language and Dataset Selection

We create a multimodal multilingual entity linking dataset in five languages, based on news articles from the M3LS dataset (Verma et al., 2023), which was originally curated for the task of multilingual, multimodal summarization. The M3LS dataset comprises news articles with associated images covering a range of topics such as politics, sports, economy, science, and technology. These articles have been curated from the British Broadcasting Corporation (BBC) over a decade and span 20 languages: 8 high-resource and 12 low-resource.

The five languages selected for our study include Hindi, Japanese, Indonesian, Tamil, and Vietnamese. Our choice is guided by several factors: *(i)* linguistic diversity, *(ii)* speaker population, and *(iii)* annotator availability.

Linguistic Diversity: The selected languages represent a broad range of language families: Austronesian (*Indonesian*), Austroasiatic (*Vietnamese*), Indo-European (*Hindi*), Dravidian (*Tamil*), and Japonic (*Japanese*). This diversity is further reflected in the various scripts used, including Devanagari (*Hindi*), Kanji/Katakana/Hiragana (*Japanese*), and Latin (*Indonesian and Vietnamese*).

Speaker Population: The chosen languages cover a wide range of the global population. Hindi is spoken by over 600 million people worldwide, Japanese by over 128 million, Indonesian by nearly 200 million (considering both native and second-language speakers), and Tamil and Vietnamese by approximately 75 million each. The inclusion of these languages in our dataset contributes to NLP research for a significant proportion of the non-English speaking global population (Blasi et al., 2022). Based on the number of articles in their dataset, M3LS classifies Tamil and Japanese as low-resource, but use Tamil and Indonesian in our dataset to advance NLP research in low-resource languages (Joshi et al., 2020).

Annotator Availability: In order to create a high-quality dataset, it is also necessary that we can identify a sufficient population of annotators who can do high-quality annotation. We utilized the Prolific Crowdsourcing Platform[2] to recruit annotators, maintaining a threshold on the languages that have a minimum of 50 eligible participants. Eligibility was determined based on participants' nationality and native language, as required by our annotation design (Table 1).

To create a balanced dataset, we selected 1,000 article titles per language from M3LS, considering factors such as budget, annotator availability, and the time required for dataset curation. More details on the selection process will be released later.

3.2 Dataset Annotation

For data annotation, we use the INCEpTION tool[3] (Klie et al., 2018), a web-based text-annotation platform designed for a range of tasks on written text, including Named Entity Recognition (NER), Part-of-Speech (POS) tagging, and Entity Linking. INCEpTION allows for the creation of projects through layers and tagsets, enabling customizable interfaces with support for both text and images. It also allows multiple collaborators to work simultaneously on the same annotation project and supports managing multiple projects concurrently. Notably, INCEpTION enables linking text to Knowledge Bases like Wikidata, which is essential for our work on entity linking, because we chose it as the knowledge base.

Nationality	#Eligible	#Selected
India	196	9
Japan	57	7
Indonesia	105	5
India	63	6
Vietnam	105	6

Table 1: Overview of eligible participants available in Prolific for the data annotation process, with a minimum threshold of 50 eligible participants.

Prolific Recruitment and Pre-screening: We used Prolific to recruit annotators based on their primary *language* and *nationality*, focusing on regions where the selected languages are most commonly spoken. To ensure the competence of the annotators, before they work on the annotation task, we implemented a two-stage process, first, a **pre-screening task** involving 10 articles in English with ground-truth labels from BBC articles, followed by the **main annotation task**. Annotators were required to achieve a minimum F1-score of 60% in pre-screening to qualify for the main annotation task. This process ensured that only qualified annotators participated in the actual annotation, where we also relied on Prolific's filters to ensure their language proficiency.

Main Annotation Task: The main annotation task involved 1,000 article titles per language, divided into 10 splits of 100 titles each to manage workload and maintain quality. Each split was annotated by **three different annotators**, and majority voting was used to determine the final entities for each mention. Annotators were provided with extensive guidelines and instructional videos to ensure consistent and accurate annotations. Further details on the number of annotators, guidelines, and links to videos used for providing instructions to annotators can be found in the Appendix.

3.3 Dataset Analysis

Entity Types: We analyze the distribution of entities based on the categories in Figure 2. We use the ''instance of'' (P31) and ''subclass of''

[2] https://www.prolific.com/.

[3] https://inception-project.github.io/.

	Hindi	Indonesian	Japanese	Tamil	Vietnamese	Overall
Number of article titles	1000	1000	1000	1000	1000	5000
Words per article title	10.48	8.85	14.99	9.19	11.99	11.10
Number of unique entities	787	765	655	572	475	2480
Number of mentions	1480	1490	1720	1254	1343	7287
Number of words per mention	1.26	1.34	1.36	1.26	1.70	1.39
Unlinked mentions*	304	201	112	456	170	1243

Table 2: Data and annotations per language, along with overall dataset statistics. Unlinked mentions (*) indicate the number of entity mentions that annotators could not link to a Wikidata page; these entities were excluded from methods.

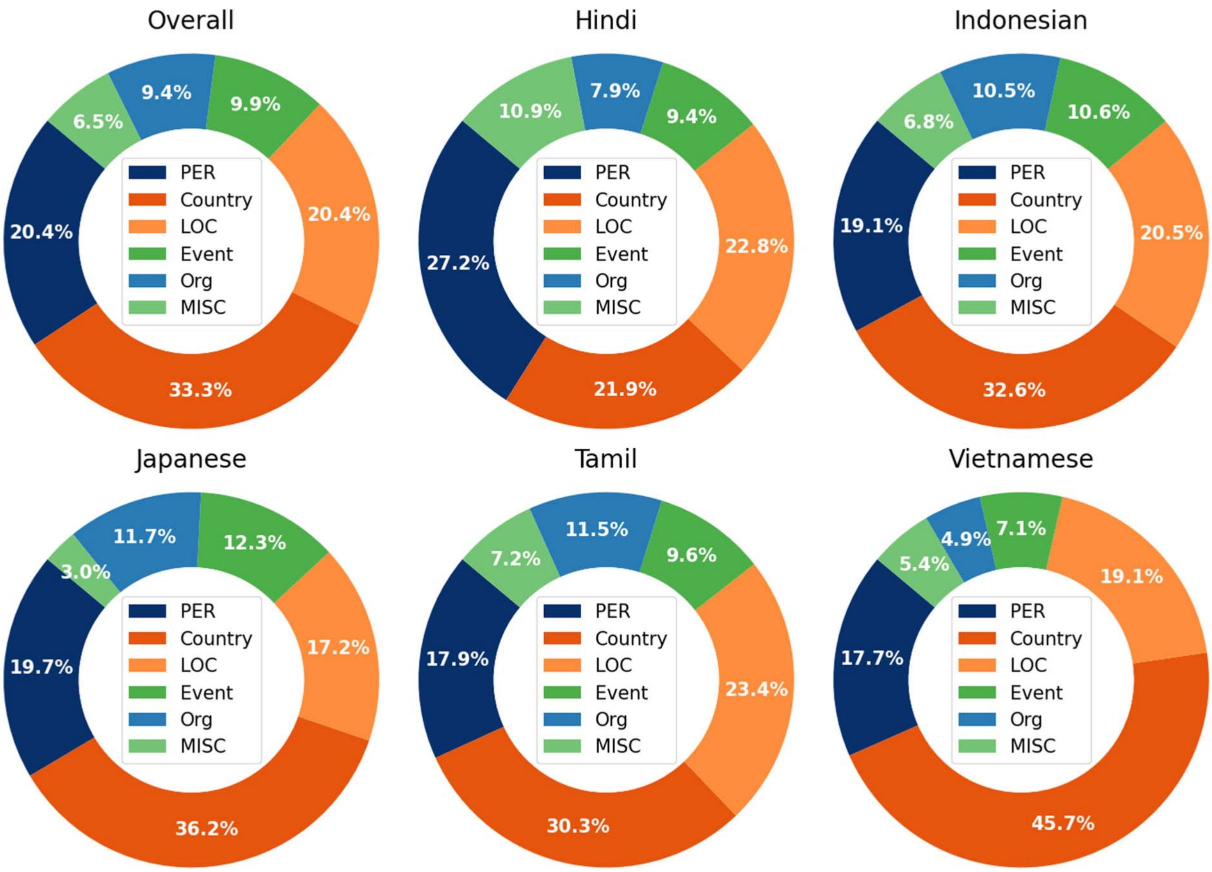

Figure 2: Category-wise distribution of entities.

(P279) properties to traverse the Wikidata hierarchy and identify the root entity, which is then mapped to either Person (PER), Country, Location (LOC), Organization (ORG), Event, or Miscellaneous (MISC).

Inter-annotator Agreement: Cohen's Kappa (McHugh, 2012) is a statistical measure used to evaluate the agreement between two annotators, taking into account the possibility of agreement occurring by chance. To calculate Cohen's Kappa for each language split between three annotators, we compare entity mentions annotated by each pair of annotators, two at a time. We first link the entity mentions marked by both annotators and focus on those where both made a selection. We then compare the Wikidata URLs chosen by the annotators to determine a match. This comparison forms the basis for calculating Cohen's Kappa, quantifying the level of agreement between annotators. The process is repeated for all annotator pairs across each language split to assess inter-annotator reliability. Table 4 shows average Cohen's Kappa values for inter-annotator agreement across five languages: Hindi, Indonesian, Japanese, Vietnamese, and Tamil. The values

MEL Datasets	KB	Multilingual	Multimodal	Languages
Mewsli (Botha et al., 2020)	Wikidata	Yes	No	ar, de, en, es, fa, ja, sr, ta, tr
Twitter (Adjali et al., 2020)	Twitter Users	No	Yes	en
Zeshel (Logeswaran et al., 2019)	Wikia (Fandom)	No	No	en
WikiDiverse (Wang et al., 2022)	Wikidata	No	Yes	en
MERLIN	**Wikidata**	**Yes**	**Yes**	**hi, id, ja, ta, vi**

Table 3: Comparison between MERLIN and other EL datasets.

Language	Average Cohen's Kappa
Hindi	0.79
Indonesian	0.81
Japanese	0.91
Vietnamese	0.88
Tamil	0.76
Overall	**0.83**

Table 4: The agreement scores reflect the consistency among annotators, with Japanese and Vietnamese showing the highest agreement, while Tamil demonstrates greater variability. Scores indicate levels of agreement as follows: ≤ 0 (no agreement), 0.01–0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and 0.81–1.00 (almost perfect).

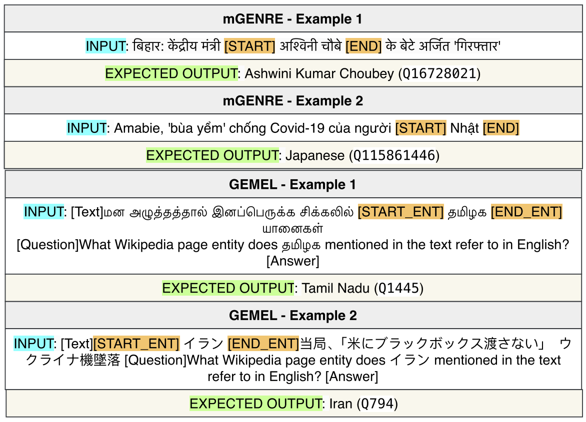

Figure 3: This figure illustrates examples of prompts used for mGENRE and GEMEL systems and how they are expected to resolve entity mentions in different languages by linking them to the correct entities.

indicate the degree of consistency in entity linking between annotators, with most values suggesting almost perfect agreement.

MERLIN vs. Other EL Datasets: Unlike other datasets that mainly emphasize text-based entity linking and are predominantly monolingual (primarily English), then MERLIN dataset combines both textual and visual data, drawing on Wikidata to cover a wide array of entity types (Figure 2) derived from news article titles. This multimodal, multilingual approach, enriched by manual curation, also incorporates the strengths of previous notable works in the entity linking domain, as illustrated in Table 3.

4 Methods

Traditional EL approaches typically employ a two-stage ''retrieve and re-rank'' strategy (Wu et al., 2020; Barba et al., 2022; Lai et al., 2022; Xu et al., 2022). The first stage involves selecting potential entities from a large KB. The second stage then re-ranks these candidates to disambiguate the final entity. However, this two-step approach can lead to compounded errors and reduced effectiveness. To address these limitations, generative approaches to entity-linking have recently been proposed, where entity names are directly generated auto-regressively. In this work, we consider two such methods: **i) mGENRE** (De Cao et al., 2022) built in the multilingual context and **ii) GEMEL** (Shi et al., 2024) built in the multimodal context. A brief overview of both methods is given below:

mGENRE: mGENRE uses a sequence-to-sequence architecture based on mBART with constrained decoding to autoregressively generate target entity names. It is a unimodal, text-only approach where it cross-encodes the text description T_m and mention m to generate the corresponding entity e in the knowledge base $\mathcal{E}$. Constrained decoding enables efficient and fast search within a large knowledge base by restricting the search space to valid Wikipedia entity names, eliminating the need for extensive vector indices. The prompts we use are shown in Figure 3. To summarize, mGENRE is a unimodal,

Pipeline	Hindi (%)		Indonesian (%)		Japanese (%)		Vietnamese (%)		Tamil (%)	
	T	T+V	T	T+V	T	T+V	T	T+V	T	T+V
mGENRE	55.72	–	**78.93**	–	71.06	–	**74.13**	–	**63.18**	–
GEMEL-Llama-2	61.1	54.17	65.96	74.29 ↑	64.59	**74.76** ↑	65.07	70.06 ↑	29.42	21.21
GEMEL-Aya-23	66.12	**67.53** ↑	70.24	75.00 ↑	69.59	70.05 ↑	69.16	69.47 ↑	47.92	37.95

Table 5: Recall@1 performance across different language pipelines. The table compares the performance of various pipelines (mGENRE, GEMEL-Llama-2, GEMEL-Aya-23 8B, and their variants without images) across five languages: Hindi, Indonesian, Japanese, Vietnamese, and Tamil. (↑) refers to cases where visual modality leads to better performance.

text-only method, which takes the article title with the entity mention as input, and directly generates the identified KB entity name as output.

GEMEL: GEMEL is adapted to **MMEL** by mapping mentions m within both visual V_m and textual T_m contexts to their corresponding entities e in a knowledge base $\mathcal{E}$. The framework keeps the vision encoder and LLM frozen, and trains a feature mapper to enable cross-modal interactions along with in-context learning demonstrations for better generations. Their work primarily focuses on multimodal EL in English and uses Llama-2 (7B) as the text encoder. We similarly train GEMEL's visual feature mapper on Wikidiverse in English (Wang et al., 2022) and test on MERLIN, zero-shot.

Given that Llama-2's (Touvron et al., 2023) pre-training data mostly consists of English tokens, we modify the architecture to include Aya-23 (8B) (Aryabumi et al., 2024) as the text encoder. Aya-23 is a multilingual LLM trained on diverse high-quality multilingual datasets, including Wikipedia. Llama-2 has limited multilingual capabilities, primarily focused on English, whereas Aya 23 is designed for strong performance across multiple languages, making it more effective in multilingual tasks.

Knowledge Base for Entity Linking: Both the baseline systems use Wikipedia titles as their knowledge bases. Constrained decoding for mGENRE is implemented through a prefix trie built on Wikipedia English titles comprising of almost 6M entities while we do not perform constrained decoding on the GEMEL methods. Since during annotation, our dataset is linked to a language-agnostic Wikidata KB, it is easy for us to preprocess the data to extract a mention's corresponding English Wikipedia titles and adapt them to these methods.

5 Results and Analysis

The performance of various entity linking methods varies significantly across different languages and models, as summarized in Table 5. To better understand these differences, we explore the following research questions.

Q1) What impact does incorporating multimodal context have in low-resource languages?

Overall Performance: The mGENRE pipeline demonstrates strong performance, particularly in Indonesian, where it achieves the highest R@1 of 78.93%. However, in other languages like Tamil, its performance is notably lower, with an R@1 of 63.18%. Both Llama-2 and Aya are not trained on Tamil data and we notice consistent low performance in Tamil by the GEMEL methods. This can also be attributed to the scripts that Tamil and even Hindi use and the lack of ability of current methods to understand contexts in these languages. The GEMEL-Llama-2 7B model shows improvements over mGENRE in languages such as Japanese, where it achieves an R@1 of 74.76%. The GEMEL - Aya-23 8B model also demonstrates competitive performance, particularly in Hindi and Tamil. In Hindi, it achieves an R@1 of 67.53%, outperforming both mGENRE and GEMEL-Llama-2 as shown in Table 5. For languages like Hindi and Tamil, the performance of Aya-23 is better than Llama-2, though both are lesser than mGENRE, which was trained extensively on all languages, otherwise both models exhibit similar performance.

Images: The Secret Sauce? We also compare our GEMEL's Llama-2 and Aya-23 variants, with and without images by introducing random noise instead of a visual prefix, to assess the significance of images in disambiguation. mGENRE and GEMEL are not directly comparable because

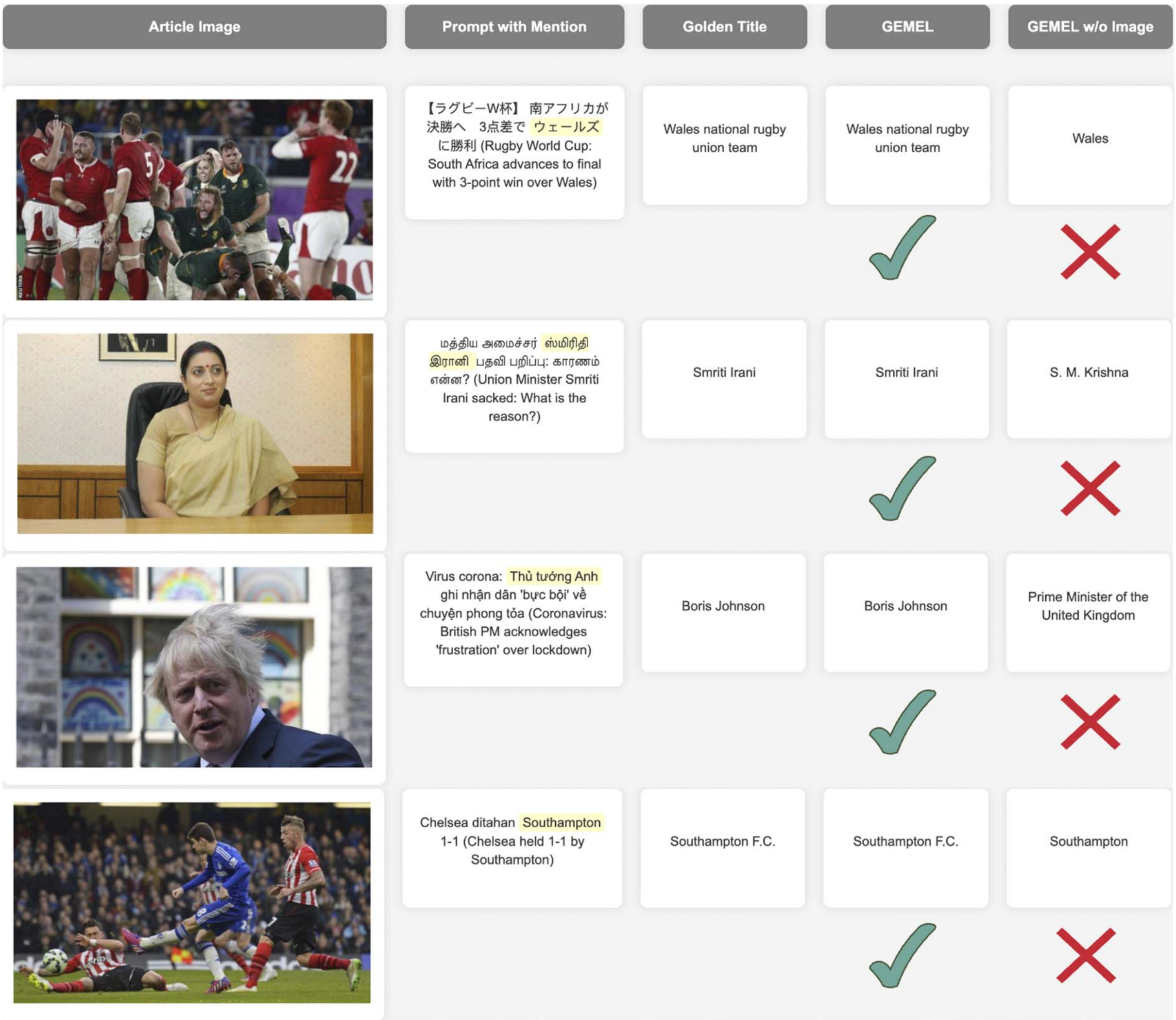

Figure 4: Impact of visual context in GEMEL. The table compares the GEMEL model's predicted wikipedia titles with and without images across four examples. Each example includes the article image, prompt with mention, golden Wikipedia title, and the predictions made by GEMEL with and without images.

both are fundamentally different architectures and mGENRE utilizes mBART as the text decoder, which is extensively multilingual and trained on over 100 languages. In contrast, Llama-2 is trained on significantly fewer multilingual datasets, and Aya-23's coverage includes only 23 languages, and lacks Tamil, a language present in MERLIN. Llama-2's performance **drops significantly** when images are removed, highlighting its strong reliance on visual data to achieve high accuracy. In contrast, Aya-23 experiences a much smaller drop without images, indicating that it is less dependent on visual input. This finding aligns with Hu et al. (2023), who emphasized the importance of multimodality in cross-lingual tasks, particularly for models that lack explicit multilingual training. Overall, we hypothesize that image support is crucial for models that aren't trained to be multilingual, like Llama-2, but its importance reduces for models trained on diverse multilingual text, like Aya-23. While we cannot include visual context to mGENRE in the current setting, we believe that the multimodal gains in GEMEL will also apply to mGENRE, especially given it's strong multilingual capabilities.

A few qualitative examples of how images help in MMEL are shown in Figure 4. In the first example, the visual context helps GEMEL correctly identify the *Wales national rugby union team*, while the text-only model incorrectly links it to the country *Wales*. In the next example, the visual context allows GEMEL to correctly identify *Smriti Irani*, whereas the text-only model incorrectly links to *S. M. Krishna*, another union minister in India. Similarly, in the third and fourth examples, the image aids in correctly identifying

	Hindi	Indonesian	Japanese	Vietnamese	Tamil
Number of Examples	112	54	50	203	77

Table 6: *Number of ambiguous mentions across languages* idenitified by filtering mentions that can be linked to multiple Wikidata IDs.

Model	Modality	Hindi (%)	Indonesian (%)	Japanese (%)	Vietnamese (%)	Tamil (%)
GEMEL-Llama-2	T+V	74.11 ↑	62.96 ↑	48.00 ↑	80.30 ↑	18.18
	T	73.21	48.15	28.00	63.55	28.57
GEMEL-Aya-23	T+V	75.00	66.67	48.00 ↑	67.00 ↑	38.96
	T	75.00	66.67	46.00	59.61	40.26

Table 7: Accuracy scores across languages for ambiguous mentions: For Hindi and Japanese, visual inputs provide marginal improvements, but they significantly help Indonesian and Vietnamese, (↑) indicates those. For Tamil, existing models struggle to get decent performance, even with visual inputs.

Southampton F.C. and *Boris Johnson*, while the text-only model fails by linking it to the city *Southampton* and the position of *Prime Minister of the United Kingdom*, respectively, which seem generic.

Q2) What kind of instances especially benefit from visual inputs?

Ambiguous Mentions: Ambiguous mentions occur when the same textual mention can refer to multiple distinct Wikidata entities. These cases are particularly challenging as they require models to rely on context—sometimes beyond the text itself—to correctly associate the mention with the appropriate entity. We identify such mentions by filtering for instances where an annotated mention can be linked to multiple Wikidata IDs (e.g., 'Apple' can refer to both the fruit (*Q89*) and the company (*Q312*) in Wikidata). Qualitative examples from our dataset for such mentions can be seen in the Appendix Figure 12. The total number of such instances identified are given in Table 6 and the accuracies of models across this subset are shown in Table 7.

From these results, it is evident that performance is consistently lower on ambiguous mentions across all languages. With visual inputs, the gains are minimal but consistent, underscoring the difficulty of these ambiguous cases in the MERLIN dataset. Japanese and Tamil, in particular, show the most significant drop in accuracy, with Llama-2 scoring just 28% and 18.18% respectively in the text-only setting. This stark drop in performance suggests that current entity linking methods, even when enhanced with multimodal context, need further refinement to better handle ambiguity, especially in lower resourced languages.

Influence of Mention Types: We report performance across different mention types such as PER (Person), LOC (Location), Country, ORG (Organization), MISC (Miscellaneous), and Event in the Appendix in Table 10. Overall, we observe that images significantly improve the accuracy for PER mentions across most languages, while ORG mentions are easier for models to resolve, with or without images. For MISC and Event mentions, we observe consistently lower accuracy scores across all models, suggesting that these types of mentions may require more contextual understanding that goes beyond the current multimodal approaches.

Q3) How does performance of both models differ across languages? From Table 5, we observe that models like GEMEL-Llama-2 and GEMEL-Aya-23 generally perform better with the inclusion of visual context, across all languages. This suggests that images significantly aid in improving accuracy, particularly in low-resource settings. We also notice that both Llama-2 and Aya-23 struggle on both Hindi and Tamil, while even though mGENRE is trained on these languages it's performance is relatively low. Except for GEMEL-Llama vs mGENRE for Japanese and GEMEL-Llama-unimodal vs mGENRE for Hindi, there are statistically significant differences between both systems ($p < 0.05$). Overall, while visual context consistently boosts performance

Model	Hindi	Indonesian	Japanese	Vietnamese	Tamil
Llama-2	77.57	85.43	81.72	84.53	80.19
Aya-23	76.68	82.83	79.69	82.13	77.87
Llama-2*	62.55	80.75	86.29	78.82	33.61
Aya-23*	71.86	80.88	84.77	74.62	42.76

Table 8: R@1 (in %) Performance Comparison on Translated Data for GEMEL methods. * represents performance of models on original untranslated dataset.

across languages, its impact is more critical in certain languages and categories, underlining the importance of multimodal approaches in enhancing entity linking accuracy, while also showing the amount of future research that needs to be explored in such languages.

Q4) To what extent does translating to English help improve performance on MERLIN? To understand the impact of language on MERLIN's performance, we translated all non-English data into English using the NLLB model (Costa-jussá et al., 2022). By comparing the English and multilingual results, we aimed to identify any gaps in understanding. This process allowed us to assess the capabilities of GEMEL-Llama-2-7B and Aya-23 8B and identify any potential biases towards the English language. The translation process led to a reduction in the number of entities due to translation inaccuracies. As a result, the evaluation was performed in a reduced entity space, focusing on those entities that were reliably translated. When translating multilingual data to English, we filtered out entities whose mentions did not appear in both the source and translated text. This filtering was necessary because translation can alter named entities through omission, transliteration differences, cultural adaptations, or structural sentence changes. By evaluating on this filtered subset, we ensured that performance differences between versions reflected the models' language capabilities rather than translation artifacts. We evaluated the models on the translated datasets to understand their performance in this limited space. A comparison of the performance of GEMEL-Llama-2-7B and Aya-23 8B across the five languages can be seen in Table 8. We observe a general improvement in performance across all languages, underscoring the disparity in understanding between English and non-English contexts in these methods. Notably, Tamil and Hindi show significant gains.

6 Related Work

The task of EL has evolved significantly, adapting to new challenges across various domains, including multilinguality, multimodality, and the utilization of diverse KBs.

Multilinguality in Entity Linking: Entity Linking in a multilingual context has seen substantial advancements, particularly with the introduction of models that can handle multiple languages and link entities across different language KBs. Early work in this area often focused on cross-lingual EL, where mentions in one language were linked to entities in another, typically English, KBs. For example, McNamee et al. (2011) and Tsai and Roth (2016b) addressed the cross-lingual EL problem, while (Sil and Florian, 2016) and Zhou et al. (2020) explored zero-shot transfer and low-resource languages.

With the advent of large language models (LLMs), approaches shifted towards more scalable and generalizable methods. Botha et al. (2020) introduced a multilingual EL model that extended bi-encoder architectures to over 100 languages, enabling linking to a language-agnostic reference KB. This was a significant departure from earlier methods that were limited to English KBs. The mGENRE (De Cao et al., 2022) model further advanced the field by using an autoregressive approach, which linked entities by generating unique tags in a multilingual setting. This method allowed for linking entities across languages without relying on monolingual KBs, marking a significant step forward in multilingual EL.

Recent works, such as BELA (Plekhanov et al., 2023), expanded the scope by incorporating dense retrieval models that support a wide range of languages, including those with limited resources. These models leverage advanced neural architectures to address challenges such as domain adaptation and zero-shot learning, which are critical for multilingual EL in diverse and dynamic environments.

Multimodality in EL: The emergence of multimodal content, particularly in social media and multimedia platforms, has driven the development of multimodal EL (MEL). (Moon et al., 2018) pioneered the MEL task, introducing a zero-shot framework that combined textual, visual, and lexical information to link entities in social media posts. However, challenges such as the availability of datasets due to privacy regulations (e.g., GDPR) and the limited scope of existing datasets have impeded progress in this area.

Subsequent efforts have focused on creating more comprehensive and challenging MEL datasets. For instance, Adjali et al. (2020) developed a Twitter-based MEL dataset, although it was limited by entity types and ambiguity. Similarly, Zhang et al. (2021) compiled a Chinese MEL dataset focusing on person entities, while Gan et al. (2021) collected MEL data from movie reviews, targeting characters and persons in the movie domain.

Recent advancements in MEL include the DWE+ model (Song et al., 2024), which addresses semantic inconsistencies by employing hierarchical contrastive learning and fine-grained feature extraction. AMELI (Yao et al., 2024) introduces attribute-aware MEL, integrating structured attributes into the disambiguation process, while DRIN (Xing et al., 2023) enhances alignment through a dynamic Graph Convolutional Network, significantly improving fine-grained alignment across multimodal data.

Multilingual, Multimodal Datasets: The intersection of multilingual and multimodal EL represents a new frontier in EL research. Recent work such as 2M-NER (Wang et al., 2024) demonstrates the potential of incorporating multimodal data into multilingual Named Entity Recognition (NER) tasks, paving the way for more comprehensive multilingual multimodal EL systems. This approach aligns closely with the objectives of our dataset, which seeks to bridge the gap between multilinguality and multimodality by providing a rich, annotated dataset that spans multiple languages and modalities.

Our contribution to this growing field is a multilingual multimodal EL dataset that supports the linking of entities across diverse languages and modalities, offering a robust benchmark for evaluating future EL models.

7 Conclusion

In this paper, we introduced MERLIN, a novel multilingual, multimodal entity linking dataset encompassing five diverse languages and scripts. The dataset, curated from BBC news articles with associated images, presents a unique opportunity to study the impact of visual information on entity linking across various languages. We explored several baseline systems, using language models like mGENRE, LLama-2 and Aya-23B, to establish a benchmark for this challenging task. Our analysis highlights the significant benefits of utilizing visual cues, particularly for disambiguating entities with ambiguous textual descriptions. However, the results also underscore the inherent difficulty of entity linking, especially in a multilingual context on low-resourced languages.

8 Limitations

Though MERLIN offers a valuable resource for multilingual multimodal entity linking research, it does have some limitations, as follows.

Different Writing Styles and Limited Genre: The current data set is drawn exclusively from BBC news articles. We tried to make it as diverse as possible, considering news types, categories, etc., but including data from domains beyond news articles would have ensured a more comprehensive evaluation of model generalization abilities to different writing styles and entity distributions. This limited genre diversity may affect the generalizability of models trained on our dataset to broader real-world applications and potentially constrains the dataset's ability to benchmark state-of-the-art approaches in more diverse scenarios.

Annotator Skill and Knowledge: Although our Cohen's Kappa values indicate strong inter-annotator agreement, the data may still contain inconsistencies due to variations in annotators' language proficiencies, general knowledge, annotation skills and inevitable human errors.

Wikipedia-based Baseline Methods: Finally, while our chosen baseline methods show good results, they link to Wikipedia page titles rather than directly to Wikidata QIDs, even though our dataset supports Wikidata linking. So they may not directly reflect the performance on a pure Wikidata QID linking task, which needs to be explored through better modeling strategies.

Acknowledgments

We would like to thank Yash Butala and Raghav Kapoor for their valuable contributions to the foundational work of MERLIN. This work was supported by a research grant from the Defense Science and Technology Agency (DSTA), Singapore.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Building a multimodal entity linking dataset from tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4285–4292. Marseille, France. European Language Resources Association. https://doi.org/10.1007/978-3-030-45439-5_31

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032v2*.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2022. ExtEnD: Extractive entity disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2478–2488, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.177

Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 179–188, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2684822.2685317

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.630

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2306–2317, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1240

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904v3*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2022. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290. https://doi.org/10.1162/tacl_a_00460

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP)*, pages 4937–4951, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.400

Xingyu Fu, Weijia Shi, Xiaodong Yu, Zian Zhao, and Dan Roth. 2020. Design challenges in low-resource cross-lingual entity linking. *arXiv preprint arXiv:2005.00692*.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: A new dataset and a baseline. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 993–1001, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3474085.3475400

Stephen Guo, Ming-Wei Chang, and Emre Kiciman. 2013. To link or not to link? A study on end-to-end tweet entity linking. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1020–1030, Atlanta, Georgia. Association for Computational Linguistics.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038v3*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.560

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Tuan Lai, Heng Ji, and ChengXiang Zhai. 2022. Improving candidate retrieval with entity profile generation for Wikidata entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3696–3711, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.292

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1335

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3): 276–282. https://doi.org/10.11613/BM.2012.031, PubMed: 23092060

Paul McNamee, James Mayfield, Dawn Lawrie, Douglas Oard, and David Doermann. 2011. Cross-language entity linking. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 255–263, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Seungwhan Moon, Leonardo Neves, and Vitor R. Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P18-1186

Mikhail Plekhanov, Nora Kassner, Kashyap Popat, Louis Martin, Simone Merello, Borislav Kozlovskii, Frédéric A. Dreyer, and Nicola Cancedda. 2023. Multilingual end to end entity linking. *arXiv preprint arXiv:2306.08896v1*.

Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7654–7665, Torino, Italia. ELRA and ICCL.

Avirup Sil and Radu Florian. 2016. One for all: Towards language independent named entity linking. In *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2255–2264, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1213

Shezheng Song, Shasha Li, Shan Zhao, Xiaopeng Li, Chengyu Wang, Jie Yu, Jun Ma, Tianwei Yan, Bin Ji, and Xiaoguang Mao. 2024. DWE+: Dual-way matching enhanced framework for multimodal entity linking. *arXiv preprint arXiv:2404.04818v1*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.263

Dongsheng Wang, Xiaoqin Feng, Zeming Liu, and Chuan Wang. 2024. 2M-NER: Contrastive learning for multilingual and multimodal ner with language and modal fusion. *Applied Intelligence*, 54(8):6252–6268. https://doi.org/10.1007/s10489-024-05490-2

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.328

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.519

Shangyu Xing, Fei Zhao, Zhen Wu, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2023. DRIN: Dynamic relation interactive network for multimodal entity linking. *Proceedings of the 31st ACM International Conference on Multimedia*. https://doi.org/10.1145/3581783.3612575

Zhenran Xu, Yulin Chen, Senbao Shi, and Baotian Hu. 2022. Enhancing entity linking with contextualized entity embeddings. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part II*, pages 228–239, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-031-17189-5_19

Barry Yao, Sijia Wang, Yu Chen, Qifan Wang, Minqian Liu, Zhiyang Xu, Licheng Yu, and Lifu Huang. 2024. Ameli: Enhancing multimodal entity linking with fine-grained attributes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2816–2834, St. Julian's, Malta. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-long.172

Li Zhang, Zhixu Li, and Qiang Yang. 2021. Attention-based multimodal entity linking with high-quality images. In *Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part II*, pages 533–548, Berlin, Heidelberg. Springer-Verlag. https://doi.org/10.1007/978-3-030-73197-7_35

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, Hua Jin, and Dacheng Tao. 2023. Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10098–10111. https://doi.org/10.1109/TKDE.2023.3250499

Shuyan Zhou, Shruti Rijhwani, John Wieting, Jaime Carbonell, and Graham Neubig. 2020. Improving candidate generation for low-resource cross-lingual entity linking. *Transactions of the Association for Computational Linguistics*, 8:109–124. https://doi.org/10.1162/tacl_a_00303

A Appendix

Entity Linking Annotation Guidelines

This section includes the complete set guidelines that were provided to annotators for the entity linking task

This annotation process/study is for a project as part of a research group at **Carnegie Mellon University**. We are looking for annotators to help us link entities/mentions in news articles to their corresponding entry in existing knowledge bases. You will be given a news article title and an associated image, which can be used in cases of ambiguity.

Terms and Concepts to Understand the Process Better

- **Mention/Entity:** These are words of interest (e.g., names of persons, locations, and companies) that are mapped from an input text to corresponding unique entities in a target knowledge base. Identifying these entities in text helps computers interpret and map the mentions to corresponding real-world objects.
- **Knowledge Base:** It is a centralized database for storing information and data. It is a structured database of knowledge that can include facts, concepts, rules, and relationships about the entities in the world. We will be using **WikiData**, and you will see entities from there in the search bar while using the tool.
- **Entity Type:** We focus on the following types: *Person, Location, Organization, Country, Events, Works, Miscellaneous*. Please refer to the *Types of Entities* section for definitions and examples of these entities.
- **Reference Materials:** The full annotation guidelines and demonstration videos are present here: guidelines, videos.

Types of Entities and Examples

These are the types of entities we are looking to link. Refer to the examples and definitions in order to get a deeper understanding of an entity. You do not have to specifically mark the entity type in the tool, but we hope the below taxonomy helps you understand what all constitutes as an entity.

Types of Entities

These are the types of entities (Table 9) we are looking to link. Refer to the examples and definitions in order to get a deeper understanding of an entity. You do not have to specifically mark the entity type in the tool, but we hope the below taxonomy helps you understand what all constitutes as an entity.

Refer to the following examples to understand what entities fall under what entity type:

1. *“iPhone 12 sales banned in France immediately, you may be surprised to hear the reason.”*
   Entities are iPhone12 (WORK) and France (COUNTRY).

2. *“Shohei Ohtani joins Dodgers for 10 years and $700 million, the highest amount in North American professional sports history.”*
   Entities are Shohei Ohtani (PER), Dodgers (ORG), and North American (LOC).

Entity Type	Description	Examples
PER: Person	Refers to a person, famous personality, leader etc. or a title	*Elon Musk, Michael Jackson, Chief Justice of Court*
ORG: Organization	Any organized body of people that is formed for a goal/purpose. This includes companies, sports-teams.	*Apple (company), Real Madrid*
COUNTRY: Country	Entity indicating a country	*Australia, China*
LOC: Location	Includes all physical locations that can be located on a map (and are not a country). Examples: continents, cities, monuments, sites, universities, buildings.	*Eiffel Tower, River Nile, Supreme Court of India, Europe*
EVENT: Events	Any entity that has duration or start/end time associated with it.	*World War 2, Mother's Day*
WORK: Works	Items associated with a specific profession, e.g., apps, products, movies, songs, etc. These could include work-specific jargons.	*iPhone-12, Titanic (the movie), Civic (car model)*
MISC: Miscellaneous	Entities that are intuitively entities but do not fall into any above categories. Example: food items, animal species.	*Tacos, Nilgai, Titanic (the ship)*

Table 9: Entity Types, Descriptions, and Examples used in the annotation process.

3. *''Why is a 'selfie' in Berlin, the capital of Germany, considered strange and selfish?''*
   Entities are Berlin (LOC) and Germany (COUNTRY).

4. *''How big a success is the new malaria vaccine R21?''*
   R21 is an entity which falls under MISC category.

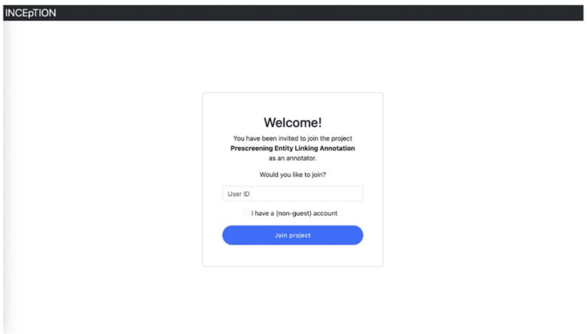

Figure 5: Login interface for the Inception annotation tool.

How to Access the website

We use the **Inception tool** for our annotation. You will be provided with the following three things:

1. Website URL
2. Assigned Documents
3. Login Instructions

Click on the URL in your preferred browser (Google Chrome is recommended). An interface like the one shown in Figure 5 will open. Enter a **User ID** to log in. Follow the naming convention suggested in the instructions. This will automatically create an account for you. **Do not** click on ''I have a (non-guest) account''. This will take you to the home page as shown in Figure 6.

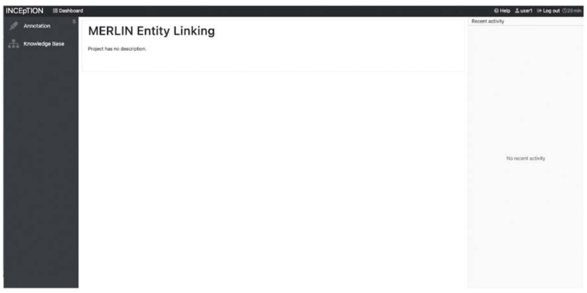

Figure 6: Inception tool home page interface.

Steps to Annotate

1. Pick the First Document

Select the first document from the **Annotation tab**. Once selected, you will see the document page interface as in (Figure 7).

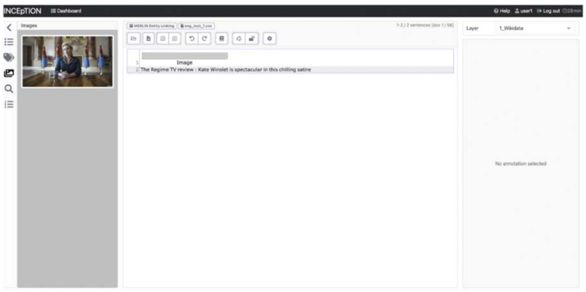

Figure 7: Document page view within the annotation tool.

2. Drag and Select Entities

There are two major components here, Image and Text (Figure 8). Here it is an Image from the TV Show ‘‘The Regime’’ and it contains the text ‘‘The Regime TV review: Kate Winslet is spectacular in this chilling satire’’

The text is interactive and you can drag and select portions of the text. This is the right way to select entities to annotate. For eg: You can drag and select mentions as shown in the image

3. Complete Annotation

Now search for the entity in the `Wikidata` search bar as shown in Figure 9 and pick the one which corresponds to the entity you think is appropriate for the text.

Note: ‘‘The Regime’’ is different from ‘‘Regime’’, so search accordingly. If needed, feel free to use google search with prompts like ‘‘The Regime Wikidata’’ or ‘‘Regime Wikidata’’ to identify the correct terms. Please refer to the **Effectively Search Wikidata A** section for more details and cases.

Make sure to also pick the Entity Type in the `NER` text box. It is a dropdown menu and the final annotation will look like Figure 11. Please refer to the Types of Entities section to understand the differences across classes.

If you cannot find the entity in the search bar AND after doing a browser/google search, select ‘‘No’’ in `Present_in_Wikidata` and leave everything blank. This is a very rare case and in most cases you should be able to find the entity (if it is frequently used or famous). Make sure you change it to ‘Yes’ if you can find one.

Do this for ALL entity spans in the same document, eg: ‘‘Kate Winslet’’ is another entity in the same document, so repeat the same procedure. Once you have completed this document, navigate to the next document/annotation through the right arrow button at the top.

Effectively Searching Wikidata Entities

In some cases, it might not be easy to find the correct entity directly from the wikidata search tab in the inception tool. Sometimes several entities with the same name would be returned which would need further digging to disambiguate and find the correct one. The search tab is also very sensitive to the textual query and it will only return the entity if there is an exact match in the name. We have listed some common cases below, please read through them thoroughly.

If you notice multiple entities in the search results, you will have to ensure that you select the correct entity. Here are some ways to identify those in cases of ambiguity:

1. **Search in English:** Search in English: If searching from the wikidata tab on the left-hand side of your tool, your search query should be in English, so for instance if your entity is ‘‘埃菲尔铁 塔’’ (English translation: Eiffel Tower), you will have to search for ‘‘Eiffel Tower’’.

2. **Exact Word Sensitivity:** Sometimes the entity’s Wikidata page name differs from your text. Example: Transliteration: ‘‘Bharat ne Sri Lanka ko 69 runo se haraya’’ (Translation: ‘‘India beats Sri Lanka by 69 runs’’), Here ‘‘Bharat / India’’ refers to the Indian cricket team. Searching for ‘‘indian cricket team wikidata’’ in Google yields the correct page: `https://www.wikidata.org/wiki/Q1143793` The title there is ‘‘India national cricket team’’, so that should be searched again in Inception (See Figure 10).

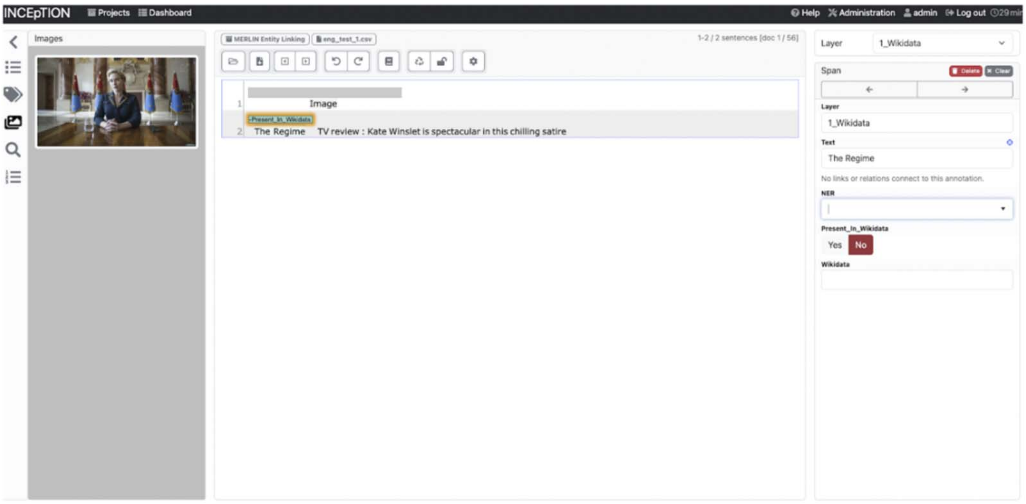

Figure 8: Selecting entity spans from text.

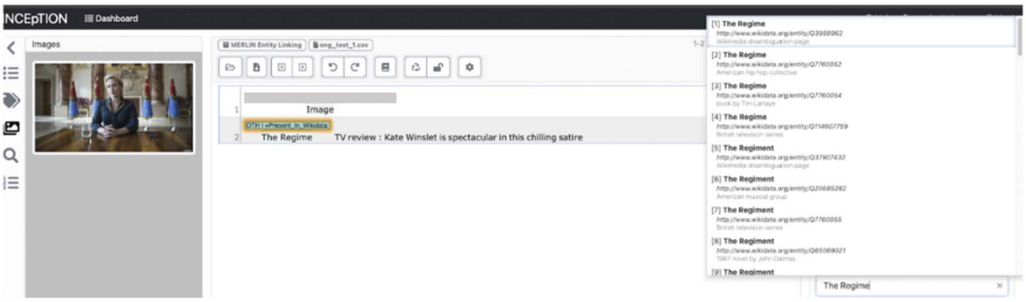

Figure 9: Searching, Disambiguating and Linking with the correct entity and type

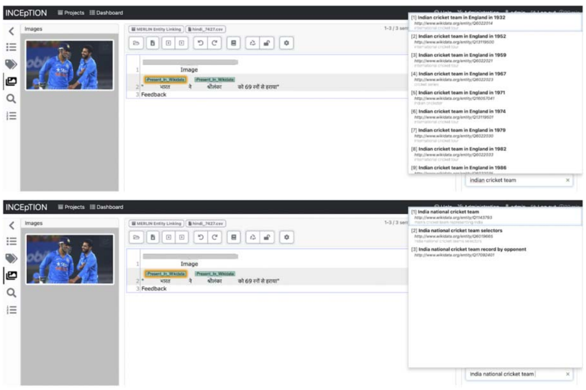

Figure 10: Example of disambiguating entities.

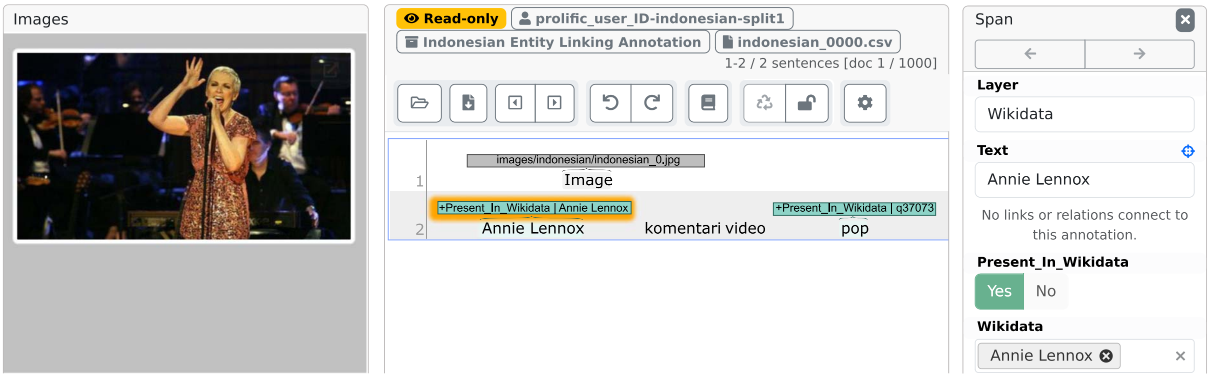

Figure 11: Entity Annotation in INCEpTION for Indonesian. The user selects and highlights text spans (e.g., ‘‘Annie Lennox’’) and assigns them to corresponding entities from a dropdown on the right side.

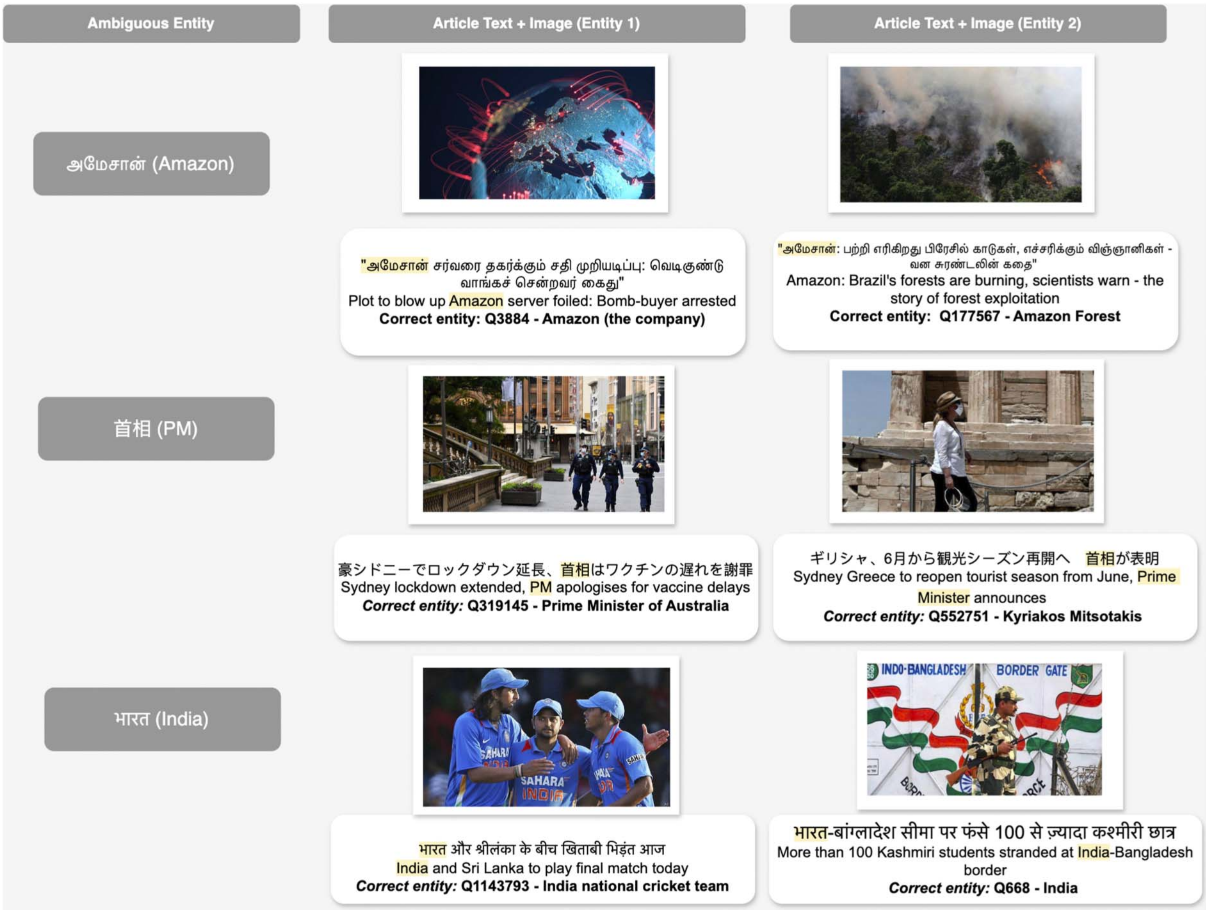

Figure 12: Ambiguous hard mentions present in MERLIN such as ‘‘Amazon,’’ ‘‘PM,’’ and ‘‘India’’ illustrate the challenge of entity disambiguation, where the same word can refer to multiple entities. For example, ‘‘Amazon’’ could refer to either the company or the rainforest, and ‘‘PM’’ might mean a Prime Minister from different countries. In these cases, combining images with textual context is crucial to accurately identify the correct entity. The visual context, such as the depiction of a forest or a tech server, narrows down the possible interpretations and we have such samples throughout the MERLIN dataset.

Model	Hindi (%)					
	PER	LOC	Country	ORG	MISC	Event
GEMEL-Llama-2	58.69	47.48	84.04	45.95 ↑	16	36.51
GEMEL-Llama-2 w/o images	62.22	57.89	92.18	43.24	28	45.24
GEMEL-Aya-23	72.80	57.28 ↑	95.77 ↑	66.67	27.33 ↑	57.14
GEMEL-Aya-23 w/o images	76.32	56.66	86.64	67.57	24.00	57.14
Model	Indonesian (%)					
	PER	LOC	Country	ORG	MISC	Event
GEMEL-Llama-2	87.54 ↑	62.81 ↑	83.86 ↑	80.54 ↑	44.44 ↑	51.77 ↑
GEMEL-Llama-2 w/o images	74.02	57.19	77.78	68.46	34.44	46.81
GEMEL-Aya-23	82.92	63.16 ↑	87.42 ↑	73.15 ↑	50.00 ↑	60.28 ↑
GEMEL-Aya-23 w/o images	83.63	56.84	79.66	72.48	48.89	50.35
Model	Japanese (%)					
	PER	LOC	Country	ORG	MISC	Event
GEMEL-Llama-2	63.31 ↑	75.59 ↑	90.69 ↑	70.65 ↑	55.77 ↑	53.55 ↑
GEMEL-Llama-2 w/o images	54.44	63.73	86.20	68.46	34.44	46.81
GEMEL-Aya-23	73.67	76.95 ↑	69.66	69.15 ↑	48.08 ↑	62.09↑
GEMEL-Aya-23 w/o images	75.15	72.54	78.33	67.66	44.23	38.66
Model	Vietnamese (%)					
	PER	LOC	Country	ORG	MISC	Event
GEMEL-Llama-2	71.01 ↑	75.88 ↑	75.57 ↑	65.15 ↑	39.73	43.16 ↑
GEMEL-Llama-2 w/o images	64.71	64.98	72.31	60.61	43.84	38.95
GEMEL-Aya-23	67.23	65.37 ↑	76.71↑	65.15	49.32 ↑	57.89 ↑
GEMEL-Aya-23 w/o images	72.27	64.59	73.13	68.18	38.36	44.21
Model	Tamil (%)					
	PER	LOC	Country	ORG	MISC	Event
GEMEL-Llama-2	24	22.11	30.79	13.19 ↑	3.33	6.61
GEMEL-Llama-2 w/o images	29.78	28.57	45.53	10.42	12.22	15.70
GEMEL-Aya-23	30.67	37.76	53.42	33.33	7.78	31.40
GEMEL-Aya-23 w/o images	42.67	42.52	66.05	49.31	16.67	35.54

Table 10: R@1 Scores (in %)for Models across Languages and Categories. This table illustrates the varying influence of visual inputs across different mention types—PER (Person), LOC (Location), Country, ORG (Organization), MISC (Miscellaneous), and Event—across five languages: Hindi, Indonesian, Japanese, Vietnamese, and Tamil. Person mentions benefit most from visual context, particularly in disambiguating textually similar entities. Country and Organization mentions are generally easier for models to resolve, with minimal gains from images. However, MISC and Event mentions remain challenging, showing lower accuracy scores across models, highlighting the need for more sophisticated approaches to handle these types of mentions effectively. The up arrow (↑) indicates an improvement in accuracy when visual inputs are included for that method.