Mahānāma: A Unique Testbed for Literary Entity Discovery and Linking

Sujoy Sarkar[1], Gourav Sarkar[1], Manoj Balaji Jagadeeshan[1], Jivnesh Sandhan[2], Amrith Krishna[3], Pawan Goyal[1],

[1]Indian Institute of Technology, Kharagpur, [2]Kyoto University, Japan, [3]BharatGen

sujoys@iitkgp.ac.in, sarkargourav12@gmail.com, manojbalaji1@gmail.com, jivnesh@i.kyoto-u.ac.jp

krishnamrith12@gmail.com, pawang.iitk@gmail.com

Abstract

High lexical variation, ambiguous references, and long-range dependencies make entity resolution in literary texts particularly challenging. We present *Mahānāma*, the first large-scale dataset for end-to-end Entity Discovery and Linking (EDL) in Sanskrit, a morphologically rich and under-resourced language. Derived from the *Mahābhārata*, the world's longest epic, the dataset comprises over 109K named entity mentions mapped to 5.5K unique entities, and is aligned with an English knowledge base to support cross-lingual linking. The complex narrative structure of *Mahānāma*, coupled with extensive name variation and ambiguity, poses significant challenges to resolution systems. Our evaluation reveals that current coreference and entity linking models struggle when evaluated on the global context of the test set. These results highlight the limitations of current approaches in resolving entities within such complex discourse. *Mahānāma* thus provides a unique benchmark for advancing entity resolution, especially in literary domains. [1]

1 Introduction

The task of Entity Discovery and Linking (EDL) must address two fundamental linguistic challenges: *variability* and *ambiguity* (Tsai et al., 2024; Rao et al., 2013). Variability refers to using different expressions to refer to the same entity, while ambiguity arises when the same expression may refer to different entities depending on the context. Successfully resolving such mentions demands a holistic understanding of discourse within or across documents (Zhou and Choi, 2018). Most studies on EDL focus on solving these challenges for named entities (NE) (Tsai et al., 2024). NEs are the central units around which document contents are organised, and accurate resolution is essential for understanding the knowledge expressed in text. Resolving named entities has been shown to enhance representation learning (Botha et al., 2020), leading to improved performance in downstream applications such as question answering (Févry et al., 2020) and knowledge extraction (Chen et al., 2021).

To address the challenges of *variability* and *ambiguity* in EDL, the task is often tackled using end-to-end Entity Linking (EL) systems, which decompose the problem into two sub-components: *mention detection* and *entity disambiguation* (Ayoola et al., 2022). Mention detection identifies spans of text that refer to entities, while entity disambiguation resolves to entries in a knowledge base (KB). A related approach is coreference resolution (CR), which clusters mentions referring to the same entity within a document, without grounding them in a KB (Lee et al., 2017). The two approaches are mutually beneficial (Arora et al., 2024; Bai et al., 2021; Durrett and Klein, 2014), and a strong cross-document coreference system could, in theory, solve EDL without a KB (Tsai et al., 2024).

However, entity resolution can be challenging in domains with high lexical variation and contextual ambiguity, particularly in literary corpora (Han et al., 2021; Bamman et al., 2020). Literary texts differ markedly from non-fictional texts like news or Wikipedia: they span long narratives, employ evolving entities and metaphorical expressions, and shift between narrative perspectives (Roesiger et al., 2018). This complexity requires deeper context modeling. Yet, most EDL research remains focused on non-literary domains such as Wikipedia (Ghaddar and Langlais, 2016; Botha et al., 2020), news (Limkonchotiwat et al., 2023), and web articles (Pradhan et al., 2012), primarily in English, leaving the challenges presented by literary texts and low-resource languages underexplored.

In this work, we present *Mahānāma*[2], a dataset

[1]*Mahānāma* is publicly available at https://github.com/sujoysarkarai/mahanama

[2]Derived from *Mahā* (Great) and Nāma (Names), signify-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24970–24984
November 4-9, 2025

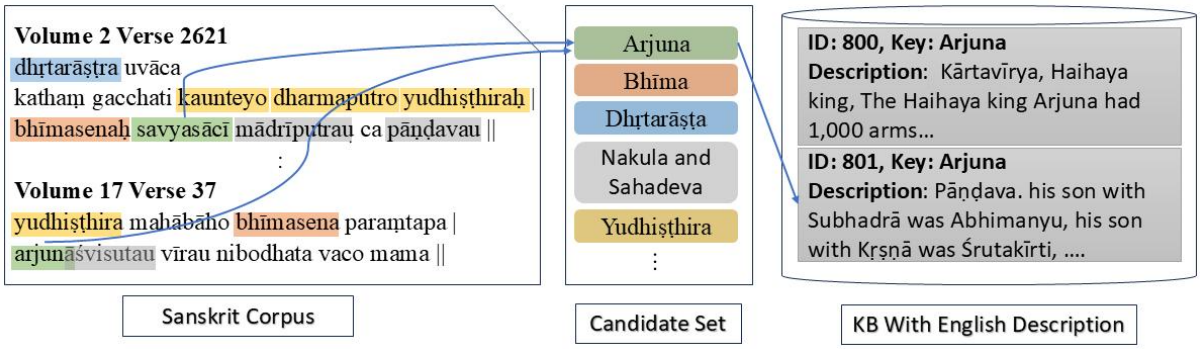

Figure 1: The figure illustrates the structure of our dataset, where name variations are highlighted in the same color. For example, *savyasācī* and *Arjuna* both refer to the same entity, *Arjuna*. Each mention is mapped to an entity, which is linked to an English knowledge base (KB) providing descriptive context. This helps distinguish between different figures sharing the same name, such as two distinct *Arjuna* entries.

constructed from the *Mahābhārata* (Dwaipāyana and Duttā, 1895), the longest epic in world literature, written in Sanskrit, a low-resource and morphologically rich language (Krishna et al., 2021). The dataset is derived from a single canonical version of the text and encompasses multiple interwoven narratives, structured in a frame-tale format (stories within a story) (Wacks, 2007). We marked 73K verses using annotation information extracted from the *"Index to the Names in Mahabharata"*(Sørensen, 1904), an existing lexicon of names in the epic. The resulting dataset includes 109K mentions spanning 5.5K entities.

Our dataset underscores the core challenges of entity resoltion. NEs in the text display significantly more variability and ambiguity than existing literary datasets. For instance, the protagonist *Arjuna* appears under 126 distinct names, while three different characters bear the same name. As shown in Figure 1, a single verse refers to *Yudhiṣṭhira* using three different names: *kaunteyo*, *dharmaputro*, and *yudhiṣṭhiraḥ*. Some entities, such as *Śiva*, have over one thousand distinct name forms. Such variation is often deeply tied to contextual and cultural cues. Characters are frequently referred to by highly context-dependent names requiring nuanced interpretation. For instance, *Arjuna* is called *Savyasācī* ("ambidextrous") to highlight his unique archery skills, and *Aindri* ("son of Indra") to indicate his divine parentage. These names may not share any lexical similarity, making their resolution especially challenging (Moosavi and Strube, 2017).

Sanskrit also introduces unique linguistic complexities. Words exhibit significant surface-form variation due to inflection and phonetic transformations at boundaries (sandhi), and its verse structure allows relatively free word order (Krishna et al., 2021; Hellwig and Nehrdich, 2018). For instance, in Example 1, the span *arjunāśvisutau* refers to three entities: *Arjuna* individually and *Nakula* and *Sahadeva* together. Here, phonetic transformation at the boundary merges *arjuna* and *aśvisutau*, altering *a* into *ā*.

$$arjuna + aśvisutau \xrightarrow{a + a = ā} arjunāśvisutau$$

Alongside the annotated corpus, we built an English KB with entity descriptions to enable cross-lingual linking between Sanskrit and English. Figure 1 shows two distinct characters named *Arjuna* from this KB, highlighting the challenge of linking across linguistically distant languages. Multilingual entity linking (MEL) resources remain scarce, with most work focusing on disambiguation rather than end-to-end processing (Botha et al., 2020).

Overall, this dataset provides a unique vantage point for analyzing EDL in settings marked by high lexical variability and ambiguity, offering a valuable resource for developing and evaluating more robust resolution systems. The following are the contributions of our work.

- We present *Mahānāma*, a large literary dataset for Entity Discovery and Linking in Sanskrit, a low-resource and morphologically rich language. The dataset contains 109K annotated mentions over 5.5K entities and captures the core challenges of EDL, namely extreme lexical variation and ambiguity. It is also accompanied by an KB with entity descriptions in English, enabling cross-lingual linking.
- We compare Mahānāma with existing literary datasets across languages and show that

ing the extensive names in the *Mahābhārata*.

it exhibits substantially higher degrees of lexical and surface-form variation and ambiguity. These characteristics pose significant challenges for current entity resolution systems.

- We conducted a manual annotation experiment involving annotators with varying familiarity with the *Mahābhārata*. Those with domain-specific background showed higher agreement with our annotations derived from the lexicon than those with only Sanskrit proficiency, suggesting that effective resolution in this dataset requires deep contextual understanding beyond basic linguistic knowledge.

- We study how variability, ambiguity, and long contextual dependencies in our dataset impact entity resolution by evaluating coreference models, including a mention-ranking baseline (Otmazgin et al., 2023) and a model designed for long texts (Guo et al., 2023). The best F1 of 51.57% highlights the difficulty of resolving context-dependent names distributed across extended narratives.

- We also assess an end-to-end multilingual entity linking model (Limkonchotiwat et al., 2023) that uses entities list, cross-lingual descriptions, and type information. While disambiguation reaches 93.27% F1 with gold mentions, overall F1 drops to 64.19% due to mention detection, showing the limits of current models in complex literary settings.

2 Related Work

The recent rise in interest in literary corpora for entity resolution has underscored challenges such as long documents, narrative complexity, and lexical variation, which are less prominent in standard datasets like AIDA (Hoffart et al., 2011) and OntoNotes (Pradhan et al., 2012).

Several corpora have been introduced to address these challenges. The DROC dataset (Krug et al., 2018) contains coreference annotations for 90 German novels with over 393K tokens. LitBank (Bamman et al., 2020) annotates the first 2,000 tokens of 100 English novels across six entity types. FantasyCoref (Han et al., 2021) covers 211 fairy tale texts. OpenBoek (van Cranenburgh and van Noord, 2022) provides 103K tokens corpus from classic Dutch novels, along with spelling normalization to account for historical language variation. KoConovel (Kim et al., 2024) focuses on 50 full-length Korean short stories, emphasizing literary resolution in underrepresented languages. Additionally, recent initiatives like CorefUD (Nedoluzhko et al., 2022) have introduced standardized multilingual coreference annotations that includes religious literary text, the Bible. Some datasets focus specifically on named entities. He et al. (2013) annotate proper names in Pride and Prejudice, while van Zundert et al. (2023) annotate character aliases in 170 Dutch novels, focusing solely on name-based identity resolution and excluding nominals and pronouns. The Friends TV show script corpus (Chen et al., 2017), in contrast, includes over 15K mentions across 46 episodes and supports both CR and EL. But, these datasets do not provide links to any external KBs.

EL and CR both begin with mention detection, but differ in how they address variation and ambiguity. EL approaches typically handle name variation through alias expansion and candidate generation (Rao et al., 2013; Özge Sevgili et al., 2022), relying on knowledge base disambiguation supported by entity types, descriptions, and alias lists (Ayoola et al., 2022; Botha et al., 2020). However, they often struggle with long-tail entities and NIL cases where no matching entry exists (Arora et al., 2024). CR models refer to it as lexical variation, encompassing named, nominal, and pronominal mentions, and address it through contextual modeling within the document. Yet their performance declines with increasing document length and lexical diversity (Joshi et al., 2019; Toshniwal et al., 2020; van Zundert et al., 2023; Arora et al., 2024). Ambiguity also knows as polysemous mentions, remains a persistent challenge for both tasks (Tsai et al., 2024). Cross-lingual EL is even less explored; Mewsli-9 (Botha et al., 2020) offers a multilingual benchmark, but is limited to newswire and centers on English as the pivot language.

To address these challenges, we present *Mahānāma*, a novel dataset for evaluating Entity Discovery and Linking in long, complex literary narratives with extensive name variation and ambiguity. It also fills a critical gap as the first large-scale resource for entity resolution in Sanskrit.

3 Dataset Creation

In this section, we present an overview of the resources used for dataset development, detail the manual efforts involved in the creation process, and describe the annotation types.

3.1 Source

Index: Our source of annotation is a book, *An Index to the Names in the Mahābhārata*, by Søren Sørensen (Sørensen, 1904). This index is a foundational reference for *Mahābhārata* studies, offering a structured catalog of names appearing in the epic. It contains approximately 12.5K primary entries, with many entries listing name variations of entities, expanding the total to around 18K names for entities. The index focuses on proper names, providing verse-level references across the 18 volumes of the *Mahābhārata*.

We utilized a digitized version of Sørensen's Index[3] (Cologne University, 2024). While the resource made the text computationally accessible, it required substantial extraction and manual correction to convert into usable annotation. Sørensen's Index provides verse references and English descriptions detailing entities and contextual roles within the *Mahābhārata*. We automatically extracted volume and verse numbers from the descriptions and retrieved all name variants linked to each entity. These clusters were then manually reviewed to ensure accurate grouping of name variants. The descriptions were used to construct a cross-lingual knowledge base (KB). Example 1 shows the descriptions of two entities in the KB.

Corpus: Multiple editions of the *Mahābhārata* exist due to its oral transmission and regional manuscript variations. Sørensen's Index refers to the Calcutta Edition (CE), which is not digitized and thus cannot be used directly. A digitized OCR version of M.N. Dutta's 1890s English translation (Dwaipāyana and Duttā, 1895), based on the CE, is available through the *Itihāsa* corpus[4] (Aralikatte et al., 2021). However, Dutta's text introduces structural modifications—merging and splitting verses, rearranging sequences, and inserting or omitting words—causing misalignment with the original. To address this, we undertook a substantial manual effort to align the 73K verses in the digitized text with the 91K verse numbers of the CE. This involved manually reading both editions and assigning CE verse numbers to the corresponding *Itihāsa* verses. Further details are provided in Appendix A. Table 1 shows an overview of the text's structure and structural difference between both editions.

[3] https://www.sanskrit-lexicon.uni-koeln.de

[4] https://github.com/rahular/itihasa

Structural Element	CE	M.N. Dutta
Volumes	18	9
Chapters	96	157
Subchapters	2110	2110
Verses	91K	73K

Table 1: Structure overview of the *Mahābhārata* (Calcutta Edition and M.N. Dutta)

Category	Entities	Mention %
Person	4.3K	91.1%
Location	0.8K	3.8%
Miscellaneous	0.4K	5.1%

Table 2: Entity distribution across categories

3.2 Annotation

Entities: The *Mahābhārata* features a vast array of entities embedded within its narrative. Sørensen's Index identifies approximately 5.5K unique entities. We manually classify these entities using the CoNLL NER tagset (Tjong Kim Sang and De Meulder, 2003) into Person, Location, and Miscellaneous categories (see Appendix B for examples). Table 2 provides distribution of these entity types.

Mentions: A mention is a linguistic expression referring to an entity in discourse (Jurafsky and Martin, 2000), including name variations and inflections. In classical Sanskrit literature, distinguishing proper names from nominals is challenging due to frequent use of compounds and derivative phrases as names, often expressing descriptions or relations (Sujoy et al., 2023), making them highly context-dependent. In our dataset, only names identified by the index are annotated as mentions; pronouns (e.g., 1, *mama* "my") and common nouns (e.g., 1, *vīrau* "two warriors") are excluded.

The corpus is unsegmented and contains multiword tokens (MWTs) (Nivre et al., 2017), where multiple words are joined together through phonological merging (sandhi) and compounding (Krishna et al., 2021). These MWTs often include more than one entity mention, with 39% of mentions in our dataset occurring within such merged forms. We annotate mention boundaries within each verse at the character level. To assist in segmenting these MWTs and identifying the start and end of inflected names, we use two tools: the Sanskrit Heritage Reader (Goyal and Huet, 2016), a lexicon-based shallow parser, and a neural network–based segmenter (Hellwig and Nehrdich, 2018). For a detailed explanation of this process, please refer to Appendix A. For example, in the

MWT *arjunāśvisutau*, two mentions are embedded: *arjuna*$_1$ and *aśvisutau*$_2$, which we annotate as:

$$arjunāśvisutau \xrightarrow{\text{Boundary}} arjunā_1, āśvisutau_2$$

Clusters and Knowledge Base: Two or more mentions referring to the same entity within a discourse are considered coreferential (Jurafsky and Martin, 2000). All occurrences of an entity name, including its name variations, are grouped into a single cluster, identified by a unique cluster ID. In addition, each cluster is linked to the KB, which provides cross-lingual descriptions in English.

Special Considerations: Our dataset explicitly marks appositive and copular mentions within the same coreference cluster, following approaches from Preco and KocoNovel (Chen et al., 2018; Kim et al., 2024). Dual and plural mentions are linked only to mentions of the same grammatical number, as per OntoNotes guidelines (Agarwal et al., 2022). Nested entities within proper names are not annotated separately to maintain consistency with prior work (Kim et al., 2024). We also include singleton entities, aligning with LitBank and Preco (Bamman et al., 2020; Chen et al., 2018), ensuring comprehensive entity coverage. Further details on these are provided in Appendix C.

3.3 Inter-Annotator Agreement

***Mahānāma* Annotation vs.**		Expert 1	Expert 2	Non-expert (Avg)
Span	κ	0.91	0.86	0.76
	F1	0.92	0.87	0.78
Span + Link	κ (All tokens)	0.89	0.81	0.69
	κ (Entity tokens)	0.80	0.67	0.53
	F1	0.80	0.68	0.56

Table 3: IAA of *Mahānāma* Annotation vs. Expert and Non-expert Annotators; κ = Cohen's Kappa

To assess annotation quality and dataset difficulty, we conducted an inter-annotator agreement study on 1,000 randomly sampled verses. Table 3 presents results for both mention detection (Span) and entity linking (Span+Link), comparing our annotations with two Sanskrit experts (both with master's degrees and expert 1 with prior experience in *Mahābhārata* studies) and a non-expert group (two students with school-level Sanskrit proficiency). We report token-level Cohen's κ for all tokens and entity tokens, and F1 scores excluding non entity tokens, as recommended by Deleger et al. (2012).

Mention detection showed high agreement across all annotator groups, with Cohen's κ indicating nearly perfect alignment with both experts (κ = 0.92, 0.86). When entity disambiguation is included, the task becomes more challenging, as reflected in a wider F1 difference between experts (0.12 for Span+Link vs. 0.05 for Span). Despite this, our annotation achieves a close to near-perfect κ of 0.80 with Expert 1 for entity linking, affirming the reliability and domain-informed accuracy of our annotations. These findings suggest that effective resolution in this dataset requires deep contextual understanding beyond basic linguistic knowledge. See Appendix D for details.

4 Dataset Analysis

This section presents our dataset's basic statistics, highlighting its unique properties by comparing it with relevant literary and non-literary entity resolution corpora (introduced in Section 2).

Basic Statistics: Our dataset contains 988,502 white space separeted tokens, making it significantly larger than other public literaray datasets for entity resolution as shown in Table 4. Additionally, our dataset is rich in NEs. Literary corpora typically have higher proportions of pronouns compared to non-literary domains(Pagel and Reiter, 2020). In our dataset, despite only NEs are marked, 10.56% of the tokens are identified as mentions, highlighting a notable entity density.

Major Entities: In literary texts, a few key entities dominate the narrative, making up most mentions (Bamman et al., 2020; Guo et al., 2023). As shown in Table 5, literary corpora typically have fewer entities than non-literary ones, with under 10% of entities contributing to over 50% of mentions. This concentration shapes the primary narrative. In our dataset, we analyze major entities at subchapter, chapter, and corpus levels. When considering the dataset as a whole, only 26 entities account for 50% of the total mentions.

Dataset	Docs	Tokens	Mentions	Entities
DROC (Lit.)	90	393K	52K	5.3K
Litbank (Lit.)	100	210K	29K	7.9K
Fantasycoref (Lit.)	214	367K	62K	6.2K
KocoNovel (Lit.)	50	178K	19K	1.4K
Openboek (Lit.)	9	103K	23.6K	8.9K
OntoNotes (Non-Lit.)	3493	1631K	194K	44K
Mewsli-9 (Non-Lit.)	58K	20M	289K	82K
Mahānāma (Lit.)	-	988K	109K (Only NE)	5.5K

Table 4: Comparison of basic statistics across literary (Lit.) and non-literary (Non-Lit.) corpora.

Lexical Variations: Our dataset shows significantly higher lexical variation in names of ma-

jor entities, with an average of 8.69 unique forms per entity at the chapter level and 124.42 at the dataset level (Table 5). For comparison datasets, we excluded only pronominal mentions and included both named and nominal forms when computing variation. Even when considering only NEs, our dataset exhibits nearly twice the variation seen in LitBank at the chapter level. At the dataset level, major entity clusters show extreme diversity, with one entity (*śiva*) appearing in up to 1,385 distinct forms. Additionally, our dataset displays exceptionally high surface-form variation due to the nature of the language.

Ambiguity: Ambiguity poses a major challenge in our dataset. As shown in Table 5, ancient literary texts such as the Bible exhibit higher ambiguity than non-literary. Notably in our datset 47% of entities share a common name, requiring context-based disambiguation essential. For example, *Janamejaya* refers to ten distinct characters. The challenge is intensified in Sanskrit, where the lack of clear markers makes it hard to distinguish proper names from common nouns (Kim et al., 2024). As seen in Figure 1, *mahābāho*(the mighty-armed) is used an adjective for *Yudhishthira*, while *mahābāhu* is also name of other distinct characters.

Spread and Burstiness: In literary texts, entities often follow a bursty pattern—long spans with few mentions punctuated by periods of intense focus (Bamman et al., 2020). Figure 2 shows the distribution of *Arjuna* across 2K subchapters, with high-frequency peaks and intermittent gaps. It also highlights a minor, overlapping entity with the same name. Resolution models must handle such burstiness and overlapping spans to accurately link mentions.

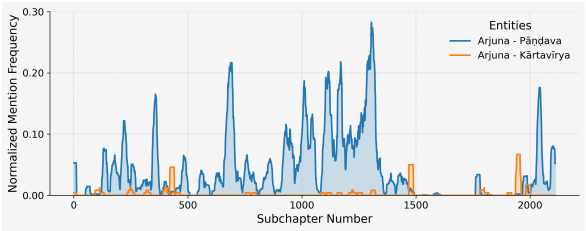

Figure 2: Mention frequency of Arjuna (Pāṇḍava) and Arjuna (Kārtavīrya) across 2K subchapters, illustrating bursty distribution and overlapping spans.

5 Experiments

We evaluate both coreference resolution (CR) and entity linking (EL) models for the task. In CR, given a document D, the goal is to cluster mentions $M = \{m_1, \ldots, m_{|M|}\}$ into entity clusters $C = \{c_1, \ldots, c_{|C|}\}$ via a function $f_{\text{CR}} : M \rightarrow C$. In EL, with a knowledge base KB of entities $E = \{e_1, \ldots, e_{|E|}\}$, the task maps mentions to entities using $f_{\text{EL}} : M \rightarrow E$. EL models rely on candidate sets and entity descriptions. We analyze the role of external knowledge and how our dataset enables studying local vs. global context in long-form narratives.

5.1 Models

As baselines, we evaluate **LingMess** (Otmazgin et al., 2023), a CR model extending the mention-ranking (MR) architecture of Lee et al. (2017), which allows us to excludes pronoun-related coreference scorers, making it suitable for our dataset. We also use **Dual Cache** (Guo et al., 2023), an entity-ranking (ER) model designed for long literary texts, which incrementally processes documents using to capture local and global entities, ideal for our dataset's structure. For multilingual entity linking, we assess **mReFiNeD** (Limkonchotiwat et al., 2023), a state-of-the-art bi-encoder model leveraging entity types and cross-lingual descriptions, ensuring robust zero-shot capabilities within an academic computational budget.

5.2 Experiment Settings

Setup: For LingMess (Otmazgin et al., 2023), we disable pronoun-related scorers due to the absence of pronoun annotations. Dual Cache (Guo et al., 2023) is configured to prevent cache misses with appropriate local and global cache sizes. Both models use Longformer-Large (Beltagy et al., 2020). mReFiNeD is trained in a multi-task setting using MuRIL (Khanuja et al., 2021) for encoding. See Appendix E for more details.

Metric: For coreference resolution, we use the standard CoNLL scorer, which reports F1 scores for **MUC**, $\mathbf{B}^3$, and $\mathbf{CEAF}_{\phi_4}$ (Moosavi and Strube, 2016). The final score is the **average F1** across these metrics. For end-to-end entity linking, we report **InKB micro-F1** with strict mention boundary matching, requiring exact matches to gold mentions. Mention detection is evaluated separately using **F1** score.

Dataset Division: EL and CR models are typically trained at the document level, each representing a single discourse. In our dataset, the entire corpus is treated as one discourse, structured as shown in Table 4. Each subchapter, averaging 468

Dataset Name	Language	Texts	Major Entities (covering 50% of mentions)					Avg. % entities with ambiguous mentions
			% of total entities	Lexical Variation (Stem)		Surface Form		
				Avg. # of variation	Max. # of variation	Avg. # of variation	Max # of variation	
DROC	German	Literary	4.99%	6.63	29	**8.26**	37	**36.23%**
Litbank	English	Literary	5.83%	4.02	20	4.19	23	10.0%
Fantasycoref	English	Literary	10.02%	**6.86**	33	7.53	34	16.0%
Openboek	Dutch	Literary	3.75%	5.26	**53**	5.50	55	25.0%
KocoNovel	Korean	Literary	18%	-	-	2.4	14	12.0%
CorefUD Proiel	Ancient Greek	Bible	9.50%	5.75	34	6.31	35	27.0%
CorefUD Proiel	Old Slavonic	Bible	10.70%	4.85	27	5.83	32	28.0%
Ontonotes	English	News, Web	24.69%	-	-	2.65	27	2.0%
Mewsli-9	11 Languages	Wikinews	4.52%	-	-	5.33	**57**	11.74%
Mahānāma (Subch.)	Sanskrit	Literary	27.56%	2.66	751	4.9	752	6.0%
Mahānāma (Ch.)	Sanskrit	Literary	5.17%	8.69	1021	27.17	1078	17.0%
Mahānāma (Total)	Sanskrit	Literary	0.46%	124.42	1385	640.58	2187	47.0%

Table 5: Comparison of dataset properties. Our dataset is analyzed at three levels—Subch (subchapter), Ch (chapter), and Total (entire dataset). For other datasets, variation includes both NE and nominal mentions, while ours is NE-only. "-" indicates low surface-form variation or unavailable stems, so lexical variation was not computed.

tokens, forms a coherent part of the Mahābhārata and serves as an independent training document. The dataset is split into 1,688 subchapters for training, 211 for development, and 211 for testing. Evaluation considers both per-subchapter performance (local) and overall test set performance (global) as a single discourse. The manually annotated 1,000 verses sampled across the text were not used for evaluation, as their scattered nature lacks the narrative context needed.

Handling Unsegmented Data: Most CR models, including the two used in our study are not designed to operate directly on unsegmented text. To address this, we adapt the Dual-Cache models to predict entity boundaries at the subtoken level as it performed better at token level. It enabled better handling of Multi-Word Tokens. This involved modifying the model code to support subtoken-level boundary prediction. For entity linking, we use character-level spans, while for coreference, entity boundaries are derived from tokenizer-generated subtokens. We evaluate both token- and subtoken-level setups to quantify their impact.

6 Results

6.1 Performance of Coreference Models

Table 6 shows CR model results, evaluated both locally (within subchapters) and globally (across the full test set) using token- and subtoken-level mention boundaries. At the token level, DualCache outperforms LingMess with an average F1 of 70.31. LingMess excels on the MUC metric (F1 79.00), which emphasizes linkage accuracy, suggesting better handling of name variations. However, it struggles with entity alignment, as seen in its low CEAFϕ_4 F1 (41.80). In contrast, DualCache performs more consistently across metrics. With subtoken-level boundary training, DualCache improves its average F1 by 4.16 points (74.46) and achieves its highest B^3 F1 (75.02), showing better mention detection and MWT handling. Globally, DualCache's CEAFϕ_4 F1 drops to 31.68, reducing its average F1 to 51.57%. While MUC remains stable, the CEAFϕ_4 drop suggests difficulty in resolving ambiguous entities across the full discourse, highlighting the need for better global resolution.

6.2 Performance of Entity Linking Model

Table 7 presents results for Entity Linking (EL), Disambiguation, and Mention Detection. mReFiNeD, applied globally, achieves an EL F1 of 64.19%, indicating potentially stronger global performance than CR models, though the scores are not directly comparable. However, its performance is limited by weak mention detection, with an F1 of 60.22%, significantly lower than DualCache (F1: 83.86%), highlighting the need to improve end-to-end models.

Ablation studies show that both cross-lingual descriptions and entity types contribute modestly to EL. Removing descriptions lowers F1 by 1.21 points, while removing types has negligible impact. This suggests that descriptions offer limited contextual benefit for resolving ambiguous entities. For entity disambiguation, which involves resolving ambiguous mentions given gold spans, mReFiNeD performs strongly with an F1 of 93.27 but relies on external resources such as a restricted set of candi-

Model	Type	Entity Boundary Marking	Eval. Level	MUC			B^3			$CEAF_{\phi_4}$			Avg.
				P	R	F1	P	R	F1	P	R	F1	F1
Lingmess	MR	Token	Local	82.30	75.90	**79.00**	76.30	67.90	71.90	74.00	29.10	41.80	64.20
Dual-Cache	ER	Token	Local	65.52	81.31	72.57	67.05	78.67	72.40	70.54	61.35	65.63	**70.30**
Dual-Cache	ER	Subtoken	Local	72.78	83.95	77.96	70.61	80.02	**75.02**	75.59	67.47	**71.30**	**74.76**
Dual-Cache	ER	Subtoken	Global	67.30	84.50	74.92	37.31	67.72	48.11	48.83	23.45	31.68	51.57

Table 6: Performance of the CR models on the test set. Model types: MR = Mention Ranking, ER = Entity Ranking

Task	Model	P	R	F1
Entity Linking	mReFiNeD	80.51	53.38	**64.19**
	w/o descriptions	79.41	52.18	62.98
	w/o entity types	80.47	53.33	64.15
Entity Disambiguation	mReFiNeD	93.30	93.24	**93.27**
	w/o descriptions	91.55	91.25	91.40
	w/o entity types	93.01	93.12	93.06
Mention Detection	mReFiNeD	63.06	57.63	60.22
	Dual-Cache	86.36	81.50	**83.86**

Table 7: Performance of models on Entity Linking, Entity Disambiguation, and Mention Detection.

Metric	Lingmess (Local)	Dual-Cache (Local)	Dual-Cache (Global)	mReFiNeD (Global)
Conf. Ent. %	10.4	3.6	7.8	2.00
Div. Ent. %	11.5	10.0	33.2	5.07
Miss. Ent. %	15.3	17.3	26.9	32.76
Miss. Ment. %	9.1	8.9	4.7	17.6
Extra Ent. %	20.0	15.2	37.7	16.5
Extra Ment. %	10.4	7.2	6.0	29.2

Table 8: Automatically identified errors percentage in predictions. **Conflated Entities:** distinct entities merged; **Divided Entity:** a single entity split into multiple; **Missing/Extra Mention/Entity:** mention/entity missing or incorrectly added. Span errors were not considered, as all spans are within single-token.

dates and their prior probabilities, underscoring the need for more self-sufficient approaches. As with EL, ablations show complementary contributions from descriptions and entity types.

7 Error Analysis

Qualitative Analysis: Both CR and EL models struggle with entity mentions in the *Mahābhārata*. The best-performing CR model fails to link lexical variations, as seen in Volume 1, Chapter 12, Subchapter 190, where the entity *draupadī* appears nine times but is split into three clusters: [*yājñasenī*, *kṛṣṇāṃ*, *yājñasenī*, *yājñasenī*]; [*pāñcālyāṃ*, *pāñcālyā*]; and [*kṛṣṇāṃ*, *draupadī*, *draupadī*], showing a tendency to group mentions based on surface similarity. It also fails to disambiguate ambiguous mentions. In Volume 7, Chapter 6, Subchapter 165, *Bhūri* (son of *Somadatta*) and *Duryodhana* (eldest son of *Dhṛtarāṣṭra*) are both referred to as *kaurava*, yet the model clusters all occurrences under a single entity.

The EL model correctly links all mentions of *draupadī* but struggles with general references. In the same document, it mistakenly links *pārtho* (plural, referring to the sons of *Pṛthā*) to *bhīma* (one of them). Similarly, in another document, *kauravaḥ* is wrongly linked to *duryodhana* instead of *bhūri*, likely due to prior probability bias. The model also struggles with mention boundary detection, especially for MWTs. These issues highlight the need for improved handling of name variations, ambiguity, context-aware resolution, and morphological richness in both approaches.

Quantitative Analysis: To assess model performance differences, we also conduct an error analysis based on the Berkeley Coreference Analyzer's error types (Kummerfeld and Klein, 2013), which categorizes errors into seven types. Table 8 presents the error distribution across models, with lower error percentage reflecting stronger performance. Refer to the Appendix F for more details.

8 Conclusion

We introduced *Mahānāma*, a large-scale Sanskrit dataset for Entity Discovery and Linking that captures challenges in literary texts, including extreme name variation, contextual ambiguity, and long-range dependencies. Derived from the *Mahābhārata*, the world's longest epic, it contains 109K mentions across 5.5K entities, annotated using a name index and linked to an English knowledge base. Evaluation of coreference and entity linking models reveals difficulty in resolving name variation and ambiguous mentions over long contexts. *Mahānāma* provides a valuable benchmark for advancing robust, context-aware entity resolution in complex literary settings.

Limitations

While *Mahānāma* makes a substantial contribution to Sanskrit entity resolution, certain limitations arise from the nature of its source material and annotation methodology. The annotations were derived automatically from a name index authored by a domain expert, which provides verse-level references without pinpointing exact name occurrences, necessitating a string-matching approach. To ensure high precision, only uniquely identifiable mentions were annotated, potentially omitting some instances. The dataset also inherits some OCR errors from the source corpora, for which no manual correction was attempted. Furthermore, the annotation focuses exclusively on named entities, excluding pronouns and common noun mentions, and is therefore not intended for comprehensive coreference resolution, though it lays the groundwork for future extensions in that direction. The definition of a "name" follows the expert author's perspective, as no standardized named entity guidelines exist for Sanskrit. While coreferential links were assigned following certain guidelines such as linking dual and plural mentions only to corresponding dual and plural entity forms. Additionally, because the dataset is based on a classical epic presented in verse format, its applicability to modern or prose texts may be limited and would require further investigation using techniques such as poetry-to-prose conversion. Since the training and test sets are drawn from the same narrative, some overlap in main entities is unavoidable, which may result in overestimation of model performance. Nonetheless, the dataset provides a valuable foundation, and future work can build upon it by exploring techniques such as data augmentation.

Ethics Statement

The annotations in this work are derived from published, copyright-free sources and a publicly available corpus. All resources utilized have been appropriately cited. The dataset, including annotations, is constructed from existing literary sources, and no explicit bias analysis has been performed. The dataset, annotations and codes is released under a CC-0 license. Annotation mapping was primarily carried out using automated methods, with experts validation conducted to ensure quality assessment and corpus alignment. Manual corpus alignment was performed by two graduate student contributors who studied Sanskrit in school, while a randomly selected set of 1000 verses was annotated by same two students and two expert with a master's degree in Sanskrit and one with a background in *Mahābhārata* studies. Annotators involved in the process were fairly compensated in accordance with standard institutional guidelines. The dataset does not contain any personal or sensitive information.

Acknowledgements

We appreciate and thank all the anonymous reviewers for their constructive feedback towards improving this work. The work was supported in part by the National Language Translation Mission (NLTM): Bhashini project by the Government of India.

AI Assistance

AI assistants such as Grammarly and ChatGPT were used in the writing process to refine textual clarity and structure.

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658, Seattle, United States. Association for Computational Linguistics.

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online. Association for Computational Linguistics.

Abhishek Arora, Emily Silcock, Melissa Dell, and Leander Heldring. 2024. Contrastive entity coreference and disambiguation for historical texts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6174–6186, Miami, Florida, USA. Association for Computational Linguistics.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint coreference resolution and character linking for multiparty conversation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 539–548, Online. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225, Vancouver, Canada. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Cologne University. 2024. Cologne digital sanskrit dictionaries, version 2.7.91. Accessed on January 30, 2024.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA Annual Symposium Proceedings*, 2012:144–153. PMID: 23304283; PMCID: PMC3540456.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Krishna Dwaipāyana and Manmatha Nāth Duttā. 1895. *Mahābhārata*. Elysium Press, Calcutta.

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.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Pawan Goyal and Gerard Huet. 2016. Design and analysis of a lean interface for sanskrit corpus annotation. *Journal of Language Modelling*, 4(2):145–182.

Qipeng Guo, Xiangkun Hu, Yue Zhang, Xipeng Qiu, and Zheng Zhang. 2023. Dual cache for long document neural coreference resolution. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15272–15285, Toronto, Canada. Association for Computational Linguistics.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1312–1320, Sofia, Bulgaria. Association for Computational Linguistics.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, Brussels, Belgium. Association for Computational Linguistics.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st edition. Prentice Hall PTR, USA.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Kyuhee Kim, Surin Lee, and Sangah Lee. 2024. Koconovel: Annotated dataset of character coreference in korean novels. *arXiv preprint arXiv:2404.01140*.

Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal. 2021. A Graph-Based Framework for Structured Prediction Tasks in Sanskrit. *Computational Linguistics*, 46(4):785–845.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, Stephan Feldhaus, and Fotis Jannidis. 2018. Description of a corpus of character references in german novels - droc [deutsches roman corpus]. *DARIAH-DE Working Papers*.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Vladimir Likic. 2008. The needleman-wunsch algorithm for sequence alignment. *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne*, pages 1–46.

Peerat Limkonchotiwat, Weiwei Cheng, Christos Christodoulopoulos, Amir Saffari, and Jens Lehmann. 2023. mReFinED: An efficient end-to-end multilingual entity linking system. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15080–15089, Singapore. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2017. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada. Association for Computational Linguistics.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Janis Pagel and Nils Reiter. 2020. GerDraCor-coref: A coreference corpus for dramatic texts in German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 55–64, Marseille, France. European Language Resources Association.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Delip Rao, Paul McNamee, and Mark Dredze. 2013. *Entity Linking: Finding Extracted Entities in a Knowledge Base*, pages 93–115. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ina Roesiger, Sarah Schulz, and Nils Reiter. 2018. Towards coreference for literary text: Analyzing domain-specific phenomena. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 129–138, Santa Fe, New Mexico. Association for Computational Linguistics.

Søren Sørensen. 1904. *An Index to the Names in the Mahabharata: With Short Explanations and a Concordance to the Bombay and Calcutta Editions and P.C. Roy's Translation*, volume 1. Williams & Norgate, London.

Leon Stassen. 1994. Typology versus mythology: The case of the zero-copula. *Nordic Journal of Linguistics*, 17(2):105–126.

Sarkar Sujoy, Amrith Krishna, and Pawan Goyal. 2023. Pre-annotation based approach for development of a Sanskrit named entity recognition dataset. In *Proceedings of the Computational Sanskrit & Digital Humanities: Selected papers presented at the 18th World Sanskrit Conference*, pages 59–70, Canberra, Australia (Online mode). Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Chen-Tse Tsai, Shyam Upadhyay, and Dan Roth. 2024. *Introduction to Entity Discovery and Linking*, pages 1–14. Springer Nature Switzerland, Cham.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: The arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Andreas van Cranenburgh and Gertjan van Noord. 2022. Openboek: A corpus of literary coreference and entities with an exploration of historical spelling normalization. *Computational Linguistics in the Netherlands Journal*, 12:235–251.

Joris van Zundert, Andreas van Cranenburgh, and Roel Smeets. 2023. Putting dutchcoref to the test: Character detection and gender dynamics in contemporary dutch novels. In *Proceedings of the Computational Humanities Research conference 2023*, pages 757–771. CEUR Workshop Proceedings (CEUR-WS.org). Computational Humanities Research Conference ; Conference date: 06-12-2023 Through 08-12-2023.

D. Wacks. 2007. *Framing Iberia: Maq?m?t and Frametale Narratives in Medieval Spain*. The Medieval and Early Modern Iberian World. Brill.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

A Annotation Mapping Process

The process of creating our dataset, illustrated in Figure 3, involved mapping the annotations provided by the "Index to the Names in the Mahabharata" to the "Itihasa Corpus". This process was divided into three main stages:

First, we extracted name variants and reference data from the index. As shown in the top-left of Figure 3, each entry in the index includes multiple variant forms of a name, with associated verse references. We manually verified and connected these name variants to ensure accurate entity resolution (e.g., airāvana and airāvata).

Second, we aligned the verse numbers from the index—originally keyed to the Calcutta edition of the Mahabharata—with those used in the Itihasa Corpus. This required manually reading and mapping verse numbers to corresponding entries in the corpus (bottom-left of the figure).

Third, we marked the occurrences of each name within the corresponding verses. This was non-trivial because the index only lists verse numbers, not the exact token positions, and the textual data is unsegmented—meaning that names may appear compounded with other words in 39% of cases.

To identify names within such tokens, we used the Sanskrit Heritage Reader (SHR), a lexicon-based shallow parser (Goyal and Huet, 2016), which could detect names in 85% of cases by examining all valid segmentations. For 12% of cases where SHR failed, we used a neural segmenter (Hellwig and Nehrdich, 2018). In the remaining 3%, where OCR errors or misspellings were present, we applied the Needleman-Wunsch approximate string matching algorithm (Likic, 2008), followed by manual correction. The final annotation, as seen at the bottom of Figure 3, links each

Type	Example with Description
Person	**Indra** – the chief of the devas, lord of rain **Aśvatthāman** – son of Droṇa and Kṛpī. **Madhusūdana** – alias of Kṛṣṇa
Location	**Kurukṣetra** – the country of the Kurus **Nīlaparvata** – a mountain **Brahmaloka** – the world of Brahman
Misc	**Sthuṇākarṇa** – name of a weapon **Mahāśaṅkha** – name of a tree **Kaumudī** - the day of full moon in the month of Kaumuda

Table 9: Examples of entity types.

token-level name occurrence back to the correct Knowledge Bases entity ID.

B Entity Types and Examples

Our annotation schema includes three coarse-grained entity types: **Person**, **Location**, and **Miscellaneous**. **Person** refers to named individuals or groups, human, personified, or divine, mentioned in the text, including relational mentions. **Location** includes named physical or conceptual places. **Miscellaneous** covers named objects, weapons, plants, any other names remaining in the index.

C Special Considerations

Apposition and Copular Mentions: Apposition occurs when two noun phrases refer to the same entity, with one providing additional information about the other. For example, in "kaunteyo dharmaputro yudhiṣṭhiraḥ" (Yudhishthira, the son of Kunti and Dharma), *kaunteyo*, *dharmaputro*, and *yudhiṣṭhiraḥ* are coreferential (Nedoluzhko et al., 2022). Copular mentions establish identity via a copula (e.g., "Yudhishthira is the son of Dharma"), but Sanskrit often omits it (zero-copula) due to its rich case system (Stassen, 1994). Following Preco (Chen et al., 2018) and KocoNovel (Kim et al., 2024), we group appositive and copular mentions into the same cluster.

Dual and Plural Mentions: Most coreference datasets assume anaphors have a single antecedent (Yu et al., 2020), with few exceptions like ARRAU (Uryupina et al., 2020). Sanskrit also features a dual grammatical number, referring specifically to two entities. For example, *mādrīputrau* and *pāṇḍavau* refer to Nakula and Sahadeva. Following OntoNotes (Agarwal et al., 2022), we mark dual and plural mentions as coreferential only with dual or plural antecedents.

Nested Mentions: Proper names are typically considered indivisible units, and any internal references within them are usually not annotated or identified (Kim et al., 2024). Following this approach, we do not explicitly mark nested mentions as coreferential. For example, in *dharmaputro* ("son of Dharma"), which refers to Yudhiṣṭhira, the nested entity *dharma* ("the god of justice") is not separately annotated.

Singletons: Singletons refer to entities with only one mention (Nedoluzhko et al., 2022). Of the 5.5K entities in our dataset, 3.1K are singletons. As our dataset provides descriptions for all entities, and recent datasets such as LitBank (Bamman et al., 2020) and Preco (Chen et al., 2018) also include singletons for coreference tasks, we choose to keep the annotation for singletons.

Unsegemeted Data: In Sanskrit, verses must adhere to one of the prescribed metrical patterns of Sanskrit prosody, which results in a relatively free word order, and words are often joined together to fit these metrical patterns (Krishna et al., 2021). This leads to phonetic transformations (Sandhi) (Hellwig and Nehrdich, 2018), merging words into continuous multi-word tokens. We keep the text unsegmented and mark entity boundaries at the character level rather than applying automatic segmentation (Hellwig and Nehrdich, 2018). 39% of mentions in our dataset consist of compounds or multi-word tokens.

1. *brahmaśiraḥ* + *arjunena* $\xrightarrow{aḥ + a = o'}$ *brahmaśiro'rjunena*

For example, in *brahmaśiro'rjunena*, *brahmaśiraḥ* ("Brahmashira weapon") and *arjunena* ("by Arjuna") merge into a single span.

D Inter Annotator Agreement Study

To carry out the inter-annotator agreement (IAA) study, three groups independently annotated a set of 1,000 randomly selected verses using an online interface that supported both span marking and entity linking to Knowledge Base. Annotators were provided with verse numbers and access to the full corpus, enabling them to refer to broader narrative context when needed. The groups included two Sanskrit experts (both with master's degrees, one with prior experience in *Mahābhārata* studies) and a non-expert group with basic Sanskrit familiarity. All annotators had general cultural exposure to the epic. Agreement was measured by comparing each group's annotations to ours using token-level Cohen's κ and F1 scores. For token-level κ, we

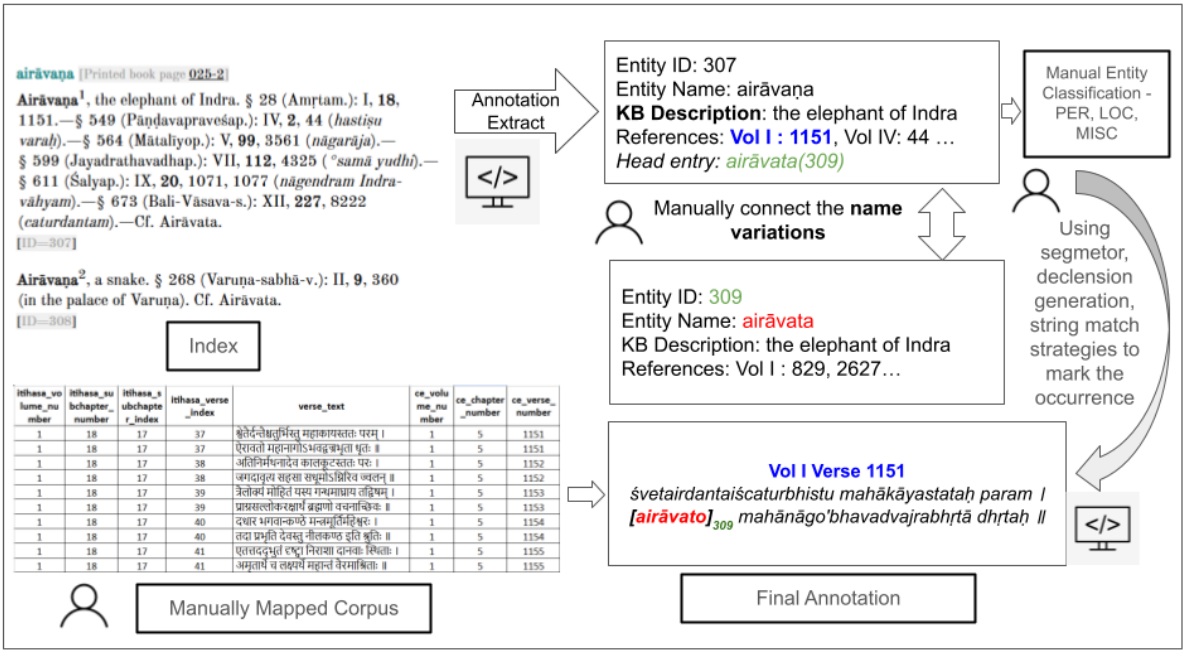

Figure 3: The annotation pipeline for mapping index entries to the Itihasa Corpus. Entity variants are manually clustered, verse references are mapped to corpus verse IDs, and final occurrences are marked using a combination of string matching strategies.

computed agreement both over all tokens and over entity tokens only (i.e., tokens part of a mention by at least one annotator). F1 scores were calculated excluding non-entity labels, following guidelines by Deleger et al. (2012).

While Cohen's κ remains a common IAA metric, it can be inflated in entity linking tasks due to token imbalance and sparse annotations. To address this, F1 scores which offers a more task-relevant view of agreement. Our annotations showed strong alignment with Expert 1 in both span detection and linking, with lower agreement observed for Expert 2 and the non-expert group—especially in the linking task. Notably, the F1 score difference between Expert 1 and Expert 2 for mention detection was modest (91 vs. 87), while the gap widened for entity linking (0.80 vs. 0.68), underscoring that disambiguation requires deeper domain understanding even among linguistically trained annotators.

E Implementation Details

We train our models using the Hugging Face library, initializing them with the Longformer-Large (Beltagy et al., 2020)[5] and MuRIL (Khanuja et al., 2021)[6] pre-trained models. Our experiments involve three models: **LingMess** (Otmazgin et al., 2023)[7], **Dual Cache** (Guo et al., 2023)[8], and **mReFiNeD** (Limkonchotiwat et al., 2023)[9].

LingMess. We disable pronoun-related scorers (`PRON-PRON-C`, `PRON-PRON-NC`, `ENT-PRON`) as our dataset lacks pronoun annotations. The model is trained for 100 epochs on an NVIDIA L40 GPU, with hyperparameters tuned for validation F1-score. Training takes approximately 18 hours.

Dual Cache. We configure the cache to prevent misses by setting the local cache (LRU) and global cache (LFU) sizes to 1000. The model is also trained for 100 epochs on an NVIDIA L40 GPU, and training requires around 34 hours.

mReFiNeD. We train mReFiNeD in a multi-task setting for mention detection, entity typing, disambiguation, and linking. We use coarse-grained tags (`PER`, `LOC`, `MISC`) and retain 30 candidates per mention, which include the gold entity, the top-ranked

[5] https://huggingface.co/allenai/longformer-large-4096

[6] https://huggingface.co/google/muril-base-cased

[7] https://github.com/shon-otmazgin/lingmess-coref

[8] https://github.com/QipengGuo/dual-cache-coref

[9] https://github.com/amazon-science/ReFinED

candidate, and random negatives. Candidate ranking uses the estimated probability $\hat{p}(e_j|m_i)$, with global priors estimated from the training corpus. Both mention and description encoders use MuRIL, a multilingual model for Indian languages. Training is done for 40 epochs on an NVIDIA A40 GPU and completes in approximately 8 hours.

We explore batch sizes of 8, 16, and 32 during hyperparameter search, while keeping other parameters aligned with the original model implementations.

F Quantitative Error Analysis

Table 8 categorizes model-specific errors using the Berkeley Coreference Analyzer framework (Kummerfeld and Klein, 2013), adapted to our single-token mention setup. The following error types were considered: *Conflated Entity*, where distinct gold entities are incorrectly merged; *Divided Entity*, where a single gold entity is erroneously split into multiple predicted clusters; *Missing Entity / Mention*, where the system fails to identify a gold entity or mention; and *Extra Entity / Mention*, where the model predicts an entity or mention that does not exist in the gold annotations.

Conflated Entity errors (e.g., 10.4% for Lingmess) occur when the model merges mentions of different entities. This aligns with the qualitative error noted where *Bhūri* and *Duryodhana* are both grouped under the common term *kaurava*, leading to incorrect entity merging due to insufficient disambiguation. These errors are highest in Lingmess and lowest in mReFiNeD, as the latter was provided with a possible alias list based on prior probability.

Divided Entity errors (e.g., 33.2% for Dual-Cache Global) reflect over-splitting of a single entity into multiple clusters. This supports our qualitative observation regarding *Draupadī*, where lexical variants like *Yājñasenī*, *Kṛṣṇāṃ*, and *Pāñcālyā* were not clustered together. These errors are highest in Dual-Cache Global, as the model struggled to connect all mentions of entities across the full test set, and lowest in mReFiNeD due to its use of a prior-based alias list.

Missing and Extra Mentions/Entities highlight the difficulty models face in detecting all valid references. For instance, **Extra Entity** errors peak at 37.7% for Dual-Cache Global due to divided entities and the model's failure to align all entities, while **Missing Entity** errors reach 32.7% for mReFiNeD due to poor mention detection in end-to-end training.

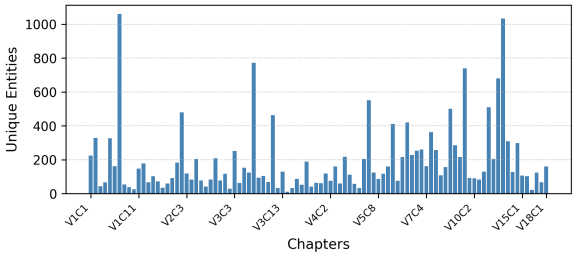

Figure 4: Distribution of unique entities per chapter in the dataset.

G Additional Dataset Statistics

The dataset exhibits an average entity density of 0.11 (i.e., roughly one entity mention every ten tokens). Entity overlap across sections is limited: the average Jaccard similarity across chapters is 0.127, indicating that only about 13% of entities, which are major entities, are shared between any two chapters. At the subchapter level, the average Jaccard 0.064.

At the chapter level, the average chain length is 5.97, while at the subchapter level it is 2.48. However, focusing on the major entities, their average chain length rises to 31.35 at the chapter level and 4.04 at the subchapter level. Thus, while many entities are ephemeral, a handful of central characters maintain long and recurrent chains that dominate the discourse. Figure 4 show the distribution of unique entities across chapters.