Mahānāma: An Epic Literary Dataset for Entity Linking and Named Entity Coreference

Anonymous ACL submission

Abstract

We present *Mahānāma*, a large-scale annotated literary dataset for entity linking and named entity coreference in Sanskrit, a low-resource and morphologically rich language. Derived from the *Mahābhārata*, the longest epic in world literature, it consists of 73K verses with 1.09M entity mentions, linked to an English knowledge base for cross-lingual resolution. Unlike previous datasets, *Mahānāma* encompasses a single long-form discourse with comprehensive entity annotations, serving as a unique testbed for end-to-end resolution tasks. The dataset poses challenges due to lexical variation, polysemous names, and long-range entity references. Experiments show that tested coreference models struggle with entity alignment across the discourse, while the entity linking model yields suboptimal performance in end-to-end linking. Cross-lingual descriptions and entity types contribute complementarily to disambiguation. *Mahānāma* provides a rich resource for studying entity linking and coreference in literary texts.

1 Introduction

Resolution tasks such as Entity Linking (EL) and Coreference Resolution (CR) are critical challenges in Natural Language Processing (NLP) that require a holistic understanding of discourse in a document or multiple documents to accurately identify and cluster entity mentions (Zhou and Choi, 2018). Entity linking grounds mentions to a knowledge base (KB) (Tsai and Roth, 2016), while coreference resolution clusters mentions referring to the same entity within a document (Lee et al., 2017). These tasks are essential for various NLP applications, such as question-answering (Févry et al., 2020) and knowledge extraction (Li et al., 2020). Entities also play a crucial role in representation learning, contributing to improved performance in downstream tasks (Botha et al., 2020).

Most research on resolution tasks has focused on 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 significant gaps in other domains such as literary texts and low-resource languages. Although CR in literary texts has recently gained attention due to their complex narratives, diverse entity mentions, and long discourse (Roesiger et al., 2018; Bamman et al., 2020; Pagel and Reiter, 2020; Han et al., 2021), EL research remains largely confined to Wikipedia-based datasets. In particular, there is a notable lack of multilingual EL (MEL) resources that support end-to-end processing (Limkonchotiwat et al., 2023), where the corpus is in one language and the KB is in another. Most existing MEL work focuses solely on entity disambiguation rather than the full pipeline of mention detection and disambiguation (Limkonchotiwat et al., 2023). Moreover, Wikipedia-based EL datasets often suffer from the problem of NIL entities (entities without a representation in a KB) and unlabeled entities (Ilievski et al., 2018; Botha et al., 2020; Limkonchotiwat et al., 2023)[1]. Likewise, CR datasets rarely account for entity references spanning multiple documents (Arora et al., 2024), limiting their applicability to broader discourse-level resolution.

In this work, we present a large-scale annotated literary dataset for end-to-end entity linking and named entity coreference in Sanskrit, a low-resource and morphologically rich language. The corpus is paired with an English knowledge base that provides entity descriptions, enabling cross-lingual linking of Sanskrit entity mentions to the English KB. Leveraging existing literary resources, we have annotated the *Mahābhārata*—the longest epic in world literature, composed in Sanskrit verse

[1] For example, in the Wikipedia page of the entity *Arjuna*, two entities, *Kichaka* and *Dīrghāyu* are mentioned but not hyperlinked, and *Dīrghāyu* lacks a Wikipedia entry

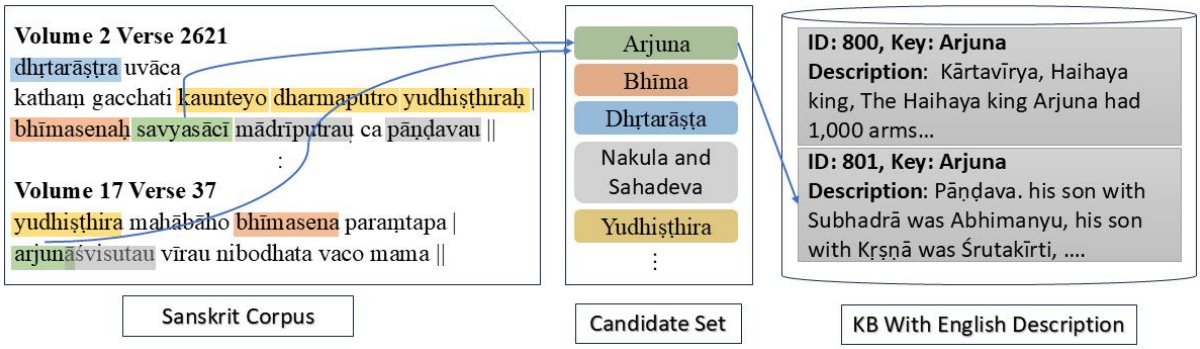

Figure 1: The figure illustrates the structure of our dataset, where coreferring mentions are highlighted in the same color. For example, *savyasācī* and *arjuna* both refer to the entity *Arjuna*. The Sanskrit corpus contains entity mentions, which are mapped to a set of candidate entities. The linked English KB provides descriptions, distinguishing different figures with the same name, such as two distinct *Arjuna* entries.

- with the occurrence of all entities. The dataset consists of 73K verses and 1.09M annotated mentions for 5.5K entities.

Our dataset, *Mahānāma*[2], derived from a single source, contains multiple stories within a unified narrative, structured in a frame-tale format (story-within-a-story)(Wacks, 2007). Recent research shows that literary texts shows markedly different characteristics (Roesiger et al., 2018), making resolution tasks more challenging (Bamman et al., 2020). Unlike non-fictional texts focused on information delivery, literature focuses on poetic descriptions and compelling storytelling. Entities in literary texts exhibit high lexical variation due to progression over long narratives (Han et al., 2021), idiomatic expressions, and paraphrasing as stylistic devices (Roesiger et al., 2018). Our dataset reflects these challenges, exhibiting long discourse and high lexical variation (Table 4). It also contains a high prevalence of polysemous or ambiguous names, a primary challenge in entity linking (Rao et al., 2013). For instance, the central character *Arjuna* has 126 distinct name variations, and 3 different characters share the same name. Figure 1 shows the English description of 2 *Arjuanas* in our KB. Similarly, the character *Yudhiṣṭhira* is referred to by 3 different names (*kaunteyo, dharmaputro, and yudhiṣṭhiraḥ*) in a single verse. Furthermore, literary texts shift between narrative spheres (Roesiger et al., 2018) and rely on implicit context, requiring deeper interpretation than structured texts like news or Wikipedia (van Cranenburgh, 2019). Moreover, our dataset presents cross-lingual linking challenge between two linguistically diverse languages, aligning with the shift toward multilingual NLP driven by advances in representation learning.

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). 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*.

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

Overall, this dataset provides a unique vantage point for analyzing resolution tasks in a linguistically rich, low-resource, and cross-lingual literary context. The dataset will be made publicly available upon acceptance. The following are the contributions of our work:

- We present a large, publicly available literaray dataset for resolution tasks in Sanskrit, a low-resource language, annotated with 1.09M mentions for 5.5K entities, categorized into three types, and paired with cross-lingual English KB containing entity descriptions.

- We evaluate both CR and EL models on our dataset. For CR, we evaluate a baseline mention-ranking model (Otmazgin et al., 2023) with an entity-ranking model (Guo et al., 2023) designed for long literary texts. While the entity-ranking model outperforms the mention-ranking model, it struggles with resolving globally distributed entity information (F1: 51.57%) in our dataset.

[2]Derived from *Mahā* (Great) and Nāma (Names), signifying the extensive names in the *Mahābhārata*.

- We evaluate a multilingual EL model (Limkonchotiwat et al., 2023) that leverages additional resources such as cross-lingual descriptions, and entity types. While disambiguation performs well (F1: 93.27%), end-to-end linking is limited (F1: 64.19%) due to poor mention detection. Ablation studies indicate that these additional resources contribute only minimal improvements to overall performance.

- We compare our dataset with other literary corpora for resolution tasks across languages, analyzing lexical variation, surface-form variation, and polysemous mentions. Our findings show a significantly higher prevalence of these phenomena in our dataset, potentially affecting model performance.

2 Dataset Creation

In this section, we present a overview of the resources used in the development of the dataset and describe the various types of annotations offered by these resources. We also provide a brief description of the dataset creation process, along with the measures implemented to ensure its quality.

2.1 Source

Index: Our source of annotation is a book, *An Index to the Names in the Mahābhārata*, by Søren Sørensen, first published in 1904 (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*. In total, it identifies around 1.2M verse references throughout the text.

We extracted the volume and verse numbers associated with each name using regular expressions. Additionally, we retrieved all name variations linked to each entity, allowing us to form clusters that group together different references to the same entity. For example, the central character *Arjuna* has 126 recorded name variations, which together contribute to approximately 6K mentions. We utilized an online version of Sørensen's Index [3], by The Cologne Sanskrit Dictionary Project

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)

(Cologne University, 2024), to extract this annotation data. In addition to verse references, Sørensen's Index provides English descriptions offering contextual details about the entities, their attributes, and their roles within the *Mahābhārata*. We extracted and cleaned these descriptions to construct a cross-lingual KB. Example 1 shows description of two entities in KB.

Corpus: Our dataset is created by marking entity annotations from Sørensen's Index onto the *Mahābhārata* corpus. Due to its long oral tradition, regional variations, and historical manuscript compilations, multiple editions of the *Mahābhārata* exist. While the core narrative remains consistent, variations occur in subplots, verse ordering, and specific word choices. Sørensen's Index references a specific edition—the Calcutta Edition—which has not been digitized. So it can not be utilized directly.

However, the Calcutta edition served as the basis for another book by M.N. Dutta (Dwaipāyana and Duttā, 1895), published in the 1890s. Dutta's book, which has been digitized using OCR as part of the *Itihāsa* corpus (Aralikatte et al., 2021)[4], introduced several modifications to the original text. These modifications include merging multiple verses, splitting single verses, altering verse sequences, and occasionally inserting or omitting words. As a result, there is a mismatch in verse numbering between the Calcutta Edition (which contains approximately 91K verses) and Dutta's version (which condenses them into 73K verses across 9 volumes). This discrepancy prevents the direct use of Index-provided verse references with the *Itihāsa* corpus. To address this, we manually aligned the Calcutta Edition with the *Itihāsa* corpus. An overview of the structure of the *Mahābhārata* is provided in Table 1.

Annotation Mapping Process In our work, we map annotations from the Index to the *Itihasa* Corpus through a structured process involving name and reference extraction, verse alignment, and

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

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

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

Table 2: Entity distribution across categories

name occurrence marking. Given the unsegmented nature of the text, we employ the Sanskrit Heritage Reader (SHR) (Goyal and Huet, 2016)[5] for lexicon-driven segmentation and a neural segmenter (Hellwig and Nehrdich, 2018) to accurately locate entity mentions within verses. For a detailed explanation of this process, please refer to Appendix B.

To quantify the quality of our dataset, we conducted an expert evaluation by manually annotating 200 randomly sampled verses and comparing them with our annotations. The evaluation yielded a mention precision of 95.38%, indicating that most identified mentions were correct, and a mention recall of 85.10%, suggesting that while a substantial number of mentions were captured, some were missed. The recall gap is primarily due to OCR errors in the *Itihāsa* corpus, some errors in annotation extracted from Index, and additional mentions present in M.N. Dutta's version. Additionally, the entity label accuracy of 98.21% confirms that the majority of mentions were correctly linked to their respective entities.

2.2 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. Table 2 provides distribution of these entity types.

Mentions: A mention is a linguistic expression that refers to an entity within a discourse (Jurafsky and Martin, 2024). In CR, mentions typically include proper names (PROP), noun phrases (NOM), and pronouns (PRON) (Bamman et al., 2020), while EL focuses mainly on PROP. However, in classical Sanskrit literature, distinguishing proper names from noun phrases is challenging due to the frequent use of compounds and derivative noun phrases as names that express descriptions or relations(Sujoy et al., 2023), which makes them highly context dependent. For instance, *Arjuna* is called *Savyasāchi* for his ambidextrous archery skills and *Aindri* as the son of *Indra*. In our dataset, only names identified by the index are annotated as mentions, while pronouns (e.g., 1, *mama* "my") and common noun mentions (e.g., 1, *vīrau* "two warriors") are excluded.

Referential Links and Knowledge Base: Two or more mentions referring to the same entity within a discourse are considered coreferential (Jurafsky and Martin, 2024). 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 an entity in the KB, which provides cross-lingual descriptions in English derived from Sørensen's Index. This enables disambiguation across mentions, enriching entity linking with additional semantic context.

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 A.

Our dataset is unsegmented and contains multi-word tokens (MWTs) (Nivre et al., 2017), where multiple words are merged due to sandhi and compounding (Krishna et al., 2021). These MWTs frequently include multiple entity mentions, with 39% of the mentions in our dataset occurring within MWTs. We identify and mark entity boundaries at the character level using the Sanskrit Heritage Reader (Goyal and Huet, 2016), a Finite State Transducer for segmentation, and a neural segmentor (Hellwig and Nehrdich, 2018). For example, consider the MWT *arjunāśvisutau*, in which the two mentions $arjuna_1$ and $aśvisutau_2$ are marked as follows:

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

3 Dataset Analysis

In this section, we present the basic statistics of our dataset, highlighting its unique properties and associated challenges. To provide context, we compare

[5] https://sanskrit.inria.fr/

it with existing literary and selected non-literary corpora for coreference resolution and multilingual entity linking corpus. For literary corpora, we reference publicly available datasets, including LitBank (Bamman et al., 2020), FantasyCoref (Han et al., 2021), OpenBoek (van Cranenburgh and van Noord, 2022), and KocoNovel (Kim et al., 2024). Additionally, we include CorefUD (Nedoluzhko et al., 2022), a collection of 21 corpora for CR across 15 languages. For MEL as datsets are mostly based on Wikipedia, we compare our dataset with only Mewsli-9 (Limkonchotiwat et al., 2023) for reference. Given the extensive number of datasets, we present statistics only for the highest-ranked CorefUD corpora alongside the other selected datasets.

3.1 Basic Statistics

Our dataset contains 988,502 white space separeted tokens, making it significantly larger than other public literaray datasets for resolution tasks as shown in Table 3. Additionally, our dataset is rich in entity mentions. Literary corpora typically have higher proportions of pronouns within coreference chains compared to non-literary domains(Pagel and Reiter, 2020). In our dataset, despite pronouns and common noun mentions not being marked, 11% 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 4, 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
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	106K	5.5K

Table 3: Comparison of basic statistics across literary (Lit.) and non-literary (Non-Lit.) corpora.

3.2 Lexical variations

Literary texts frequently employ lexical variation and paraphrasing as stylistic devices (Roesiger et al., 2018), leading to entities being referenced by multiple expressions (Han et al., 2021). This poses challenges for CR and EL tasks, especially when nominal phrases lack head matching (Moosavi and Strube, 2017). For example, *savyasācī* in example 1 refers to *Arjuna* without a direct lexical match. In entity linking, this is termed "name variation" when only proper names are involved (Agarwal et al., 2022).

Our dataset, despite excluding pronoun and common noun mentions, exhibits significant lexical variation, with an average of 8.69 unique name variations per entity at the chapter level and 124.42 at the dataset level (Table 4). This is nearly twice the variation observed in LitBank at the chapter level. At the dataset level, major entity clusters demonstrate extreme diversity, with one entity (*śiva*) having up to 1,385 distinct name variations. For other datasets, we excluded only pronouns and considered all mention types when calculating variation.

3.3 Polysemous Names

Polysemous name, where a single name refers to multiple entities(Chen et al., 2021), is a significant challenge in our dataset. For example, the name *"Janamejaya"* corresponds to ten distinct characters. This ambiguity challenge is widely recognized in EL(Rao et al., 2013). As shown in Table 4, literary texts, particularly those in ancient languages, exhibit higher polysemy than non-literary ones. In our dataset, 47% of entities share a common name, making context-based disambiguation essential.

3.4 Spread and Burstiness

In literary texts, entities that spread over a long text range often exhibit a bursty pattern, characterized by periods of sparse or no mentions followed by intense focus(Bamman et al., 2020). Figure 2 illustrates the spread and frequency distribution of the major entity, *Arjuna*, across 2K subchapters, displaying high-frequency peaks interspersed with periods of low or no mentions. Additionally, Figure 1 shows a minor but long-range entity, also named *Arjuna*, whose span overlaps with the primary *Arjuna* entity. Resolution models must learn to connect mentions while accounting for the bursty distribution and overlapping spread typical of entities in literary texts.

Dataset Name	Language	Texts	Major Entities (covering 50% of mentions)					Avg. % entities with polysemous mentions
			% of total entities	Lexical Variation (Stem)		Surface Form		
				Avg. # of variation	Max. # of variation	Avg. # of variation	Max # of variation	
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 4: Comparison of dataset properties. Our dataset is analyzed at three levels—Subch (subchapter), Ch (chapter), and Total (entire dataset). "-" indicates low surface-form variation or stem not available, so lexical variation was not calculated. For other datasets, only pronouns were excluded, while all mention types were considered for variation.

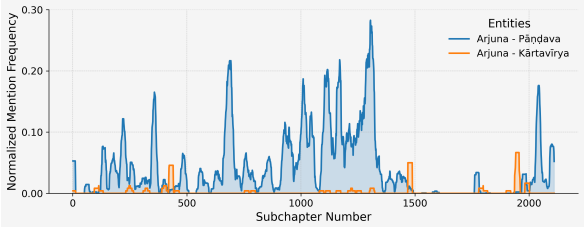

Figure 2: Mention frequency of Arjuna (Pāṇḍava) and Arjuna (Kārtavīrya) across 2K subchapters, illustrating bursty distribution and overlapping spans.

3.5 Language Specific Challenges

Languages like Sanskrit pose unique challenges for resolution task due to the lack of specific markers that clearly differentiate between common noun phrases and proper names (Kim et al., 2024). For instance, in Table 1, the term "*mahābāho*" (the mighty-armed), used as an epithet for *Yudhishthira*, is not a name. However, in certain contexts, *mahābāhu* refers to two sons of the character *Dhṛtarāṣṭra*, demonstrating the potential ambiguity in name usage. Additionally, our dataset exhibits exceptionally high surface-form variation in entity names compared to other corpora owing to the nature of the language (Table 4).

4 Experiments

We evaluate CR and EL models using our dataset. In CR, given a document D, the task is to cluster mentions $M = \{m_1, \ldots, m_{|M|}\}$ into equivalence classes $C = \{c_1, \ldots, c_{|C|}\}$, where each c_i represents a unique entity, using a function $f_{CR} : M \rightarrow C$. In EL, given a knowledge base KB with entities $E = \{e_1, \ldots, e_{|E|}\}$, the task maps mentions M to entities E via a function $f_{EL} : M \rightarrow E$. EL models leverage candidate sets and entity descriptions. We analyze how external knowledge aids in resolving entities, while our dataset's single-discourse nature enables local vs. global context resolution in long narratives.

4.1 Models

As baselines, we evaluate **LingMess** (Otmazgin et al., 2023), a coreference resolution 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 which is lacking pronoun annotations. We also use **Dual Cache** (Guo et al., 2023), an entity-ranking (ER) model designed for long literary texts, which incrementally processes documents using L-cache and G-cache 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 and efficiency within an academic computational budget.

4.2 Experiment Settings

Setup: For LingMess (Otmazgin et al., 2023), we disable pronoun-related scorers (`PRON-PRON-C`, `PRON-PRON-NC`, `ENT-PRON`) as our dataset lacks pronoun annotations. For Dual Cache (Guo et al., 2023), we analyze cache misses and set cache sizes to `LRU` (local) and `LFU` (global) at 1000, preventing misses. Both models use Longformer-Large (Beltagy et al., 2020) as the encoder. For mReFiNeD, we

train 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, including 1 gold, top-ranked, and random negatives. Candidate ranking is based on $\hat{p}(e_j|m_i)$, with global priors estimated from the corpus. We use MuRIL (Khanuja et al., 2021), a multilingual language model specifically built for Indian languages, as both the mention and description encoder. For other hyperparameters please refer to C.

Metric: For coreference resolution, we use the standard CoNLL scorer, which evaluates performance based on **MUC**, $\mathbf{B}^3$, and $\mathbf{CEAF}_{\phi_4}$ metrics (Moosavi and Strube, 2016). The final score is computed as the **averaged F1** across these three metrics. For end-to-end entity linking and disambiguation, we report **InKB micro-F1** with strict mention boundary matching, requiring predicted mentions to exactly match gold-standard annotations. Additionally, for mention detection, we evaluate performance using the **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 3. Each subchapter, averaging 468 tokens, forms a coherent part of the Mahabharata story 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.

Data Processing Our dataset is unsegmented, which can affect token-level models. We address this by using subtokens from the tokenizer, marking entity boundaries at the subtoken level for coreference models, while character-level marking suffices for entity linking model. This allows training coreference models on unsegmented data. We evaluate both token-level and subtoken-level annotations to assess their impact on models performance (See A).

5 Results

5.1 Coreference Resolution

Table 5 presents the results for CR models, evaluated both locally (within subchapters) and globally (across the entire test set) using token- and subtoken-level mention boundaries. At the token level, DualCache outperforms LingMess, achieving an average F1 score of 70.31. However, LingMess excels on the MUC metric (F1 79.00), which emphasizes linkage accuracy between mentions, suggesting it captures coreference links more effectively. But LingMess struggles with entity grouping and alignment, as evidenced by its low CEAFϕ_4 F1 score (41.80). In contrast, DualCache performs more consistently across metrics. With subtoken-level boundary marking, DualCache achieves its highest B^3 F1 score (75.02), demonstrating that subtoken boundaries improve mention detection, particularly for MWT mentions.

However, when evaluated globally, DualCache's CEAFϕ_4 F1 drops to 31.68, reducing its average F1 to 51.57%. While the stable MUC score suggests preserved linkage accuracy, the sharp decline in CEAFϕ_4 indicates that the model struggles to maintain consistent entity alignment across the entire discourse, highlighting the need for better integration of global context.

5.2 Entity Linking

Table 6 presents results for Entity Linking, Disambiguation, and Mention Detection. mReFiNeD achieves an EL F1 of 64.19%, suggesting potentially better global performance than CR models, though the scores are not directly comparable. However, its effectiveness is constrained by poor mention detection, as indicated by its standalone mention detection F1 of 60.22%, which is significantly lower than Dual-Cache (F1: 83.86%), highlighting the need for improvements in end-to-end models.

Ablation studies reveal that both cross-lingual descriptions and entity types contribute modestly to entity linking performance. Removing descriptions results in an F1 drop of 1.21 points, while removing entity types leads to an insignificant drop. This suggests that cross-lingual descriptions provide some contextual information for disambiguation, though their impact is limited.

For entity disambiguation—the task of resolving ambiguous mentions to their correct entities given gold mentions—the mReFiNeD model achieves a strong F1 score of 93.27. Similar to entity linking, ablation studies show that both entity types and cross-lingual descriptions contribute complementarily. However, the model remains dependent on external resources such as restricted candidate sets and entity priors, underscoring the need for more self-sufficient approaches.

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 5: 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 6: Performance of models on Entity Linking, Entity Disambiguation, and Mention Detection.

Metric	Dual-Cache (Local)	Dual-Cache (Global)	mReFiNeD (Global)
Gold Ent.	2900	782	782
Pred. Ent.	3327	1628	553
Conf. Ent.	135	61	52
Div. Ent.	348	260	80
Miss. Ment.	626	319	542
Extra Ment.	606	412	46
Miss. Ent.	497	211	289
Extra Ent.	551	373	40

Table 7: Automatically identified errors in predictions. **Conflated Entities:** distinct entities merged; **Divided Entity:** a single entity split into multiple; **Missing/Extra Mention:** mention missing or incorrectly added; **Missing/Extra Entity:** entity missing or incorrectly introduced. Span errors were not considered, as all spans are single-token.

6 Error Analysis

Qualitative Analysis: Both CR and EL models struggle with entity mentions in the *Mahābhārata*. The best-performing coreference model fails to link lexical variations, as seen in Volume 1, Chapter 12, Subchapter 190, where *draupadī* appears nine times but is split into three clusters: [*yājñasenī*, *kṛṣṇāṃ*, *yājñasenī*, *yājñasenī*]; [*pāñcālyāṃ*, *pāñcālyā*]; and [*kṛṣṇāṃ*, *draupadī*, *draupadī*]. The model also fails to disambiguate polysemous 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 the mentions of *Draupadī* but struggles with general names. In the same document, it mistakenly connects *pārtho* (plural mention referring to the sons of *Pṛthā*) to *Bhīma* one of the sons of *Pṛthā*). Similarly, in the second document, *kauravaḥ* is incorrectly linked to *Duryodhana* instead of *Bhūri*, likely due to frequency bias. The model also struggles with mention boundary detection, especially for MWTs. These challenges highlight the need for improved handling of name variations, polysemy, context-aware resolution, and morphologically rich languages in both models.

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 7 presents the error distribution across models, with fewer errors reflecting stronger performance.

7 Conclusion

Our work introduces *Mahānāma*, a comprehensive dataset for entity linking and named entity coreference in Sanskrit, addressing key challenges in literary discourse, including lexical variation, polysemous names, and long-range entity references. Evaluations highlight the limitations of existing models in maintaining entity alignment across discourse and the reliance of EL models on external resources. This underscores the need for improved methods that better integrate global context and cross-lingual information. We hope this dataset serves as a valuable resource for advancing research in resolution tasks for the low-resourced Sanskrit language and, more broadly, in literary domains.

Limitations

The annotation information is derived from a book authored by an expert. Although mention information was not verified against current coreference or entity linking guidelines, coreferential links were assigned following certain guidelines, such as linking dual and plural mentions only to corresponding dual and plural entity forms. Our dataset focuses solely on named entities, excluding pronoun and common noun mentions, limiting its applicability for full resolution tasks, including pronominal coreference. Additionally, some OCR errors are present in the selected corpora, but no corrections attempt have been made. The index used for annotation provides only verse numbers without specifying name occurrences within verses, requiring a string-matching approach where only uniquely identifiable mentions were marked, potentially leading to some omissions. While quality checks were conducted on randomly selected verses, annotation errors may still exist. As future work, we plan to validate the entire test set through expert review to enhance dataset reliability.

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. Both the dataset and annotations will be released under a CC-0 license. Annotation mapping was primarily carried out using automated methods, with expert 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 200 verses was annotated by an expert with a master's degree in Sanskrit and a background in *Mahābhārata* studies. Experts involved in the process were fairly compensated in accordance with standard institutional guidelines. The dataset does not contain any personal or sensitive information.

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.

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.

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.

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.

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.

Filip Ilievski, Piek Vossen, and Stefan Schlobach. 2018. Systematic study of long tail phenomena in entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 664–674, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released August 20, 2024.

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.

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.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. 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.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

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. 2019. A dutch coreference resolution system with an evaluation on literary fiction. *Computational Linguistics in the Netherlands Journal*, 9:27–54.

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.

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.

A 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.

B Annotation Mapping Process

The process of creating our dataset involved mapping the annotations provided by "An Index to the Names in the Mahabharata" to the Itihasa Corpus. This process was divided into three main tasks. First, we extracted name variations and mention information from the index, ensuring accuracy by having a subject expert manually review the entities' names and their associated variations. Second, we aligned the verse numbers from the Calcutta edition text with the Itihasa Corpus, as the index only provides verse references. This manual alignment enabled us to utilize the verse numbers from the index with the Itihasa corpus. Third, we marked the occurrences of names within each verse. Since the index specifies only the verse number without the exact position of the name, and due to the unsegmented nature of the data (where names may be joined with other words or names in 39% of cases), this task required additional tools.

To locate names within tokens, we employed the Sanskrit Heritage Reader (SHR), a lexicon-driven shallow parser (Goyal and Huet, 2016), which successfully handled 85% of cases by searching for names across all lexically valid segmentations of a token. For cases where SHR failed (12%), we utilized a neural segmenter(Hellwig and Nehrdich, 2018). In a small fraction of cases (3%), where OCR errors were present, we applied the Needleman-Wunsch approximate string match algorithm(Likic, 2008), followed by manual verification, to approximately match names. The resulting dataset is available on [github_link].

C Implementation Details

We train our models using the Hugging Face library, initializing them with the Longformer-Large (Beltagy et al., 2020)[6] and MuRIL (Khanuja et al., 2021)[7] pre-trained models. Our experiments involve three models: **LingMess** (Otmazgin et al., 2023)[8], **Dual Cache** (Guo et al., 2023)[9], and **mReFiNeD** (Limkonchotiwat et al., 2023)[10], with hyperparameters optimized to maximize the F1-score on

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

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

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

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

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

the validation set. We explore batch sizes of 8, 16, and 32, while other hyperparameters are adopted from the original model papers. After selecting the best hyperparameter configurations, we train Lingmess and Dual-Cache for 100 epochs each, while mReFiNeD is trained for 40 epochs. The training is conducted on NVIDIA L40 GPUs for Lingmess and Dual-Cache, whereas mReFiNeD is trained on an NVIDIA A40 GPU. The total training time is 18 hours for Lingmess, 34 hours for Dual-Cache, and 8 hours for mReFiNeD.