Lightweight Multilingual Coreference Resolution without LLMs @CRAC2026

Sobha Lalitha Devi, Aashik Ali S, Gopinath P,
Vijay Sundar Ram R and Pattabhi RK Rao
AU-KBC Research Centre
MIT Campus of Anna University
Chennai, India
sobha@au-kbc.org

Abstract

This paper describes our multilingual coreference system developed for the CRAC 2026 unconstrained track. We introduce a unified, single-model architecture based on Conditional Random Fields (CRFs) that supports 20 languages. Notably, our approach achieves multilingual resolution without the use of large language models (LLMs) or pretrained weights. In contrast to resource-intensive neural methods, the proposed model is efficient, and suitable for deployment on standard hardware (CPUs). It uses linguistic and contextual features to capture coreference relations across languages with diverse syntactic and morphological properties. Model training was conducted using the official data distributions released for the CRAC 2026 shared task. This methodology provides a robust, scalable solution for multilingual NLP, demonstrating high utility within resource-constrained environments. The results highlight that feature-driven structured models remain effective for complex cross-lingual tasks. The performance on test data is similar to the results obtained for the development data.

1 Introduction

This paper describes our submission to the CRAC 2026 shared task on Multilingual Coreference Resolution (Unconstrained track). Our system utilizes the provided datasets, which comprise diverse languages annotated according to the CorefUD collection guidelines.

Coreference resolution involves clustering mentions within a text that refer to the same underlying entity. Our work is driven by two key objectives: the development of a unified, language-independent model capable of universal application across all target languages, and the implementation of a system optimized for standard CPU architectures. By avoiding heavy hardware requirements and language-specific silos, we provide a highly accessible solution for multilingual processing.

In this task we have chosen the "Coreference from Scratch data" provided by the task organizers.

The primary objective of this work is to develop a unified multilingual model for coreference resolution. Our approach utilizes a sequential pipeline where mention detection is performed first, followed by the identification of coreference relations to link these mentions into cohesive chains.

2 Related Work

The history of coreference resolution (CR) has evolved through three distinct paradigms: rule-based systems, supervised machine learning, and modern end-to-end neural architectures. While global research has matured rapidly, the development of CR for Indian languages remains a critical, albeit underserved, frontier.

Early CR research (1960s–1970s) relied on handcrafted syntactic and semantic rules. Key milestones include Hobbs (1978) and Mitkov (1997; 1998), who formalized anaphora resolution algorithms. By the early 2000s, the field shifted toward supervised learning, sparked by Soon et al. (2001) and a series of influential shared tasks, such as SemEval-2010 (2010) and CoNLL-2011/2012 (2011; 2012). These tasks provided the first large-scale, unrestricted coreference datasets in languages like English, Chinese, and Arabic.

The neural revolution introduced end-to-end models that eliminated the need for separate mention detection and clustering stages. Notable advancements include Lee et al. (2017), who introduced the first end-to-end system using Bi-LSTMs and head-finding attention; Joshi et al. (2020), who leveraged SpanBERT to significantly improve contextual representations; and Kantor and Globerson (2019), who introduced "Entity Equalization" to share information across all mentions in a chain. Other parallel developments include fast end-to-end resolution frameworks for morphologically di-

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 199–204

verse languages like Korean (2020) and multi-task dialogue frameworks (2021).

Despite global progress, Indian languages characterized by complex morphology and scarcity of annotated data have historically lagged. Early efforts focused on specific linguistic challenges, such as the ICON-2011 NLP Tools Contest (2011), which established benchmarks for pronominal resolution in languages like Hindi, Bengali, and Tamil. Subsequent milestones focused on hybrid and multilingual approaches:

- **Hybrid Systems:** Pattabhi et al. (2007) combined Vector Space Models (VSM) with rule-based heuristics for English and Tamil. Vijay Sundar Ram and Sobha (2012) and Vijay Sundar Ram (2019) utilized Tree-CRFs and morphological features to handle the structural complexities of Tamil.

- **Social Media Focus:** The FIRE 2020 and 2022 SocAnaRes tasks addressed the unique difficulties of informal, multilingual social media text in Bengali, Hindi, Malayalam, and Tamil (2020; 2022).

- **Modern Neural Approaches:** Recent research has embraced Transformer-based architectures. Sobha Lalitha Devi et al. (2023; 2024) explored AdapterFusion and XLM-RoBERTa models, demonstrating that incorporating specific linguistic features into neural frameworks significantly boosts performance in Hindi, Malayalam, and Tamil.

Most recently, to combat this English-centric bias, the MCR-IL 2025 task at RANLP (2025) focused on Hindi and Dravidian languages (Kannada, Malayalam, Tamil, and Telugu), pushing the boundaries of multilingual coreference chain generation. From early handcrafted rules to modern cross-lingual transfers, the journey of coreference resolution reflects a broader shift toward inclusive, data-driven NLP that is finally beginning to address the rich linguistic diversity of the Indian subcontinent.

3 System Description

We utilize the 27-dataset 'Coreference from Scratch' collection to train a unified model supporting 20 languages. Our approach prioritizes computational and temporal efficiency by employing a CRF-based architecture that avoids the high overhead of modern transformer-based models, allowing for training and inference on CPU systems. The system architecture is bifurcated into two primary modules: Mention Detection (identifying relevant textual spans) and Coreference Resolution (establishing relational links to form complete coreference chains).

3.1 Stage 1 — Feature extraction from Conll-U to an Intermediate Format

To accommodate the requirements of the CRF-based architecture, the original CoNLL-U files (linguistically rich representation) are preprocessed into a simplified, tab-delimited structure. This transformation extracts specific linguistic attributes into a columnar format, structured as follows:

```
doc_id | sentence_number | token_number | word |
    UPOS | XPOS | chunk | lemma+morphology |
      dependency_relation | coreference_tag
```

Key features extracted are:

- **Chunk Tags:** Derived from part-of-speech heads. Nouns, pronouns, proper nouns, determiners, and adjectives all become NP (noun phrase); verbs become VP; prepositions become PP. Each chunk span is marked B-NP (beginning) or I-NP (inside).

- **Flat / Compound Names:** Names like "Peter Pett" or "Mrs. Marjoribanks" are kept together as one continuous I-NP span using dependency relations like *flat*, *flat:name*, and *nmod:desc*.

- **Empty Nodes:** Pro-dropped zero subjects (written as 23.1 in CoNLL-U) are preserved with a special `_E` suffix on their token number.

- **Document Boundaries:** Marked `# newdoc` in CoNLL-U and written as `__DOCBREAK__` lines so each document is processed independently.

- Existing gold Entity=coreference annotations are read from the `MISC` field and converted to BIO tags (B-3, I-3, etc.), though for prediction purposes these are replaced by the model's output.

3.2 Stage 2 — Mention Extraction

There are 4 PASS stages in mention extraction:

- **PASS 1 — Definite Descriptions (Highest Priority):** The code looks for phrases that

match a pre-loaded gazetteer list of "definite descriptions" — such as "the president", "Mr.", "Mrs.", "Dr." etc. collected from the given training data for all languages. When there is an exact match with the phrases in the gazetteer list, the phrase in the input is considered as a candidate mention. The mention boundaries are then dynamically extended to encompass any immediately following proper nouns; for instance, if there is 'Mrs.' And 'Pett', then it would be expanded to 'Mrs. Pett'.

Additionally, titles associated with the dependency relation *nmod:desc* (e.g., 'Dr. Rajesh Kumar') are automatically extracted as complete entity spans.

- **PASS 2 — NP Chunks:** To handle multi-entity noun phrases, our logic identifies NP boundaries via standard BIO tags. However, if a single span contains a definite description alongside a proper noun, it is split into two separate mentions. We justify this dual-mention extraction by the fact that both components (e.g., 'the chief minister' and 'Rajesh Kumar') may serve as independent antecedents or referents in subsequent discourse. Mentions are classified as `PROPER` (proper nouns/names), `COMMON` (common nouns), `DEMONSTRATIVE` (demonstrative phrases like this, that, etc.), or `DEFINITE` (definite descriptions).

- **PASS 3 — Standalone Pronouns:** Any pronoun (UPOS `PRON`, or XPOS `PRP` / `PRP$`) not already captured in the first two passes is added as its own mention of type `PRONOUN`. This includes possessive pronouns and pronominal adverbs (words like "there" or "here" that have `PronType=` in their morphology).

- **PASS 4 — Zero / Pro-drop Pronouns:** Empty nodes (like 23.1) with UPOS `PRON` are added as `ZERO` type mentions. These are the implied subjects common in pro-drop languages like Tamil. They have no surface word form but carry coreference meaning.

3.3 Stage 3 — Feature Extraction and Pairwise Prediction

Now that there is a list of all mentions, the system needs to decide which pairs of mentions refer to the same entity.

Feature Extraction (per mention): For each mention, the code extracts: Word form and lemma/root of the head word; Morphological features (gender, number, person e.g., `Gender=Masc|Number=Sing|Person=3`); Case marker (e.g., nominative, accusative etc. for morphologically rich languages); Clause information (the dependency relation of the mention's head); Sentence ID and token position (start and end token indices); and an Is-demonstrative flag.

Pair Creation & CRF Prediction: During the relation identification phase, each identified mention is paired with every other mention within a 15-sentence contextual window. These candidate pairs are then processed by a Conditional Random Field (CRF) model, which performs a binary classification to determine if the mentions corefer ('yes') or not ('no'). To ensure high precision, the system utilizes a confidence-based decision threshold (defaulting to 70%). If the model predicts a coreferent relationship with a probability below this threshold, the link is discarded and reclassified as 'no,' effectively filtering out low-confidence matches.

3.4 Stage 4 — Cluster Building and Writing Back to Conll-U

Cluster Building (Union-Find): All accepted "yes" pairs become coreference links. These links are grouped into clusters using a Union-Find (disjoint set) algorithm. This means: if A links to B, and B links to C, then A, B, and C all end up in the same cluster — they all refer to the same entity. Each cluster gets a unique numeric entity ID (e.g., e1, e2, etc.). These IDs are globally unique across all documents in a multi-document file.

Annotation Building: For each cluster, the code determines which CoNLL-U tokens belong to each mention's span (start token and end token). It then creates `Entity=` annotation strings:

`(e3-1)` represents a single token which means entity 3 and there is only 1 head token.

`(e3-2)` represents a multi token scenario where which means entity 3 and 2 represents the number of head tokens in the span. The span can be 2 or more tokens.

The syntactic head within a span is found by looking at the dependency tree. The token whose HEAD point outside the span is the syntactic head.

Writing the Output to CoNLL-U format: A two-pass process writes the final output file. Pass 1 reads the original CoNLL-U to build a dependency-

head lookup table (needed to compute head offsets for multi-token spans). Pass 2 re-writes every line of the original CoNLL-U, but: a) strips any existing `Entity=` annotation from the `MISC` column; b) injects the newly predicted `Entity=` annotation where appropriate; c) adds a `# global.Entity = eid-etype-head-other` header line at the start of each document.

4 Results

We evaluated the system's output using the organizers' evaluation script. Performance results across the mini-test set are detailed in Table 1. As we can observe, the average score is less compared to methods using LLMs and other pre-trained models. The system can be further improved by capturing more linguistic rules/patterns which are language specific.

Dataset	Accuracy Score
fr_litbankfr	28.55
ca_ancora	38.80
cs_pcedt	25.77
cs_pdt	36.77
cs_pdtsc	37.91
cu_proiel	34.80
de_potsdamcc	29.77
en_fantasycoref	41.35
en_gum	44.23
en_litbank	34.41
es_ancora	37.35
fr_ancor	37.38
fr_democrat	35.50
grc_proiel	43.36
hbo_ptnk	48.19
hi_hdtb	46.19
hu_korkor	29.51
hu_szegedkoref	31.65
ko_ecmt	21.62
la_coreflat	16.17
lt_lcc	27.97
nl_openboek	34.51
no_bokmaalnarc	42.93
no_nynorsknarc	39.65
pl_pcc	36.25
ru_rucor	33.72
tr_itcc	37.21
Average	**35.24**

Table 1: Evaluation Results for mini-test data in Unconstrained Task

A specialized study focusing on Hindi and English yielded an accuracy of 69.76% when utilizing an expanded set of language-specific features (Devi et al., 2014). Since then, coreference resolution methodologies for low-resource Indian languages such as Hindi, Tamil and Malayalam have significantly shifted from rule-based and shallow feature engineering toward deep neural architectures and multi-task learning frameworks. Table 2 details the cross-era performance comparisons for Hindi and English CR systems, showcasing how contemporary models utilize multilingual embeddings and adapter modules to give better performance without the use of LLMs.

Despite this performance, we noted a significant limitation in recall due to the absence of features for embedded mention links. These findings highlight a specific area for future optimization, particularly in resolving complex, nested coreference chains.

5 Conclusion

We successfully submitted our test runs for the CRAC 2026 Multilingual Coreference Resolution task, showcasing a lightweight CRF-based system capable of supporting 20 languages. By prioritizing a single-model approach that operates efficiently on standard CPU architectures, we have addressed the need for accessible NLP tools in settings where high-end computational power is unavailable. This submission validates the efficacy of traditional statistical modeling paired with robust linguistic features in the modern era of multilingual processing.

Acknowledgments

The authors would like to thank the organizers of CRAC 2026 for providing the datasets and the opportunity to participate in this shared task. This work was supported by the Ministry of Electronics and Information Technology (MeitY), Government of India.

References

Sobha Lalitha Devi. 2020. 'SocAnaRes-IL20: Anaphora resolution from social media text in indian languages @ fire 2020 - an overview'. In *Forum for Information Retrieval and Evaluation-2020*, IDRBT, Hyderabad, India.

Sobha Lalitha Devi. 2022. Anaphora resolution from social media text in indian languages (SocAnaRes-IL): 2nd edition-overview. In *FIRE '22: Forum for Information Retrieval Evaluation*.

Sobha Lalitha Devi, Vijay Sundar Ram, and Pattabhi R. K. Rao. 2014. A generic anaphora resolution engine for indian languages. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1824–1833.

Sobha Lalitha Devi, Vijay Sundar Ram, and Pattabhi R. K. Rao. 2024. End to end multilingual coreference

System / Study	Core Methodology	Performance Metric
Lalitha Devi (2014) (Devi et al., 2014)	Generic anaphora resolution engine using expanded language features	69.76% (F1 Accuracy)
Lalitha Devi (2020) (Devi, 2020)	Shared task framework on microblog (Twitter) datasets for messy texts	Benchmark data (SocAnaRes-IL)
Lalitha Devi (2023) (Devi et al., 2023)	End-to-end neural framework using AdapterFusion multi-task learning	68.30% (F1 Accuracy)

Table 2: Coreference Resolution system performance results for English, Hindi of our AU-KBC systems

resolution for indian languages. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 256–259.

Sobha Lalitha Devi, Vijay Sundar Ram, and Pattabhi R. K. Rao. 2025. "Multilingual Coreference Resolution - Indian Languages (MCR-IL2025) @ RANLP 2025 - an overview". In *Proceedings of the 1st Shared Task on Multilingual Coreference Resolution – Indian Languages (MCR-IL2025), at RANLP 2025*, Bulgaria. Association for Computational Linguistics.

Sobha Lalitha Devi, Vijay Sundar Ram R., and Pattabhi R. K. Rao. 2023. Coreference resolution using AdapterFusion-based multi-task learning. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON 2023)*.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):339–352.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677.

Kenton Lee, Luheng Rest 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.

Shujie Li, Hayashi Kobayashi, and Vincent Ng. 2021. The CODI-CRAC 2021 shared task on anaphora, bridging, and discourse deixis resolution in dialogue: A cross-team analysis. In *Proceedings of the CODI-CRAC 2021 Shared Task on Anaphora, Bridging, and Discourse Deixis in Dialogue*, pages 71–95.

Ruslan Mitkov. 1997. Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches. In *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, pages 14–21, Madrid, Spain.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '98/ACL '98)*, pages 869–875, Montreal, Canada.

Chaehun Park, Jaehun Shin, Sanghyu Park, Jong-Hoon Lim, and Changki Lee. 2020. Fast end-to-end coreference resolution for korean. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2610–2624.

R. K. Rao Pattabhi, L. Sobha, and Amit Bagga. 2007. Multilingual cross-document co-referencing. In *Proceedings of 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, pages 115–119, Portugal.

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.

Sameer Pradhan, Lance Ramshaw, Mitchell prestige Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27.

Vijay Sundar Ram and Sobha Lalitha Devi. 2012. Coreference resolution using Tree-CRF. In *Computational Linguistics and Intelligent Text Processing, Springer LNCS*, volume 7181, pages 285–296.

Vijay Sundar Ram R. 2019. *Resolution of Coreference Chains in Tamil*. Ph.D. thesis, Anna University Chennai.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Veronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 1–8, Uppsala, Sweden.

L. Sobha, Sivaji Bandyopadhyay, Vijay Sundar Ram R., and A. Akilandeswari. 2011. NLP tool contest @ICON2011 on anaphora resolution in indian languages. In *Proceedings of ICON 2011*.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.