Few-Shot Coreference Resolution with Semantic Difficulty Metrics and In-Context Learning

Nguyen Xuan Phuc, Dang Van Thin
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23521213@gm.uit.edu.vn
thindv@uit.edu.vn

Abstract

This paper presents our submission to the CRAC 2025 Shared Task on Multilingual Coreference Resolution in the LLM track. We propose a prompt-based few-shot coreference resolution system where the final inference is performed by Grok-3 using in-context learning. The core of our methodology is a difficulty-aware sample selection pipeline that leverages Gemini Flash 2.0 to compute semantic difficulty metrics, including mention dissimilarity and pronoun ambiguity. By identifying and selecting the most challenging training samples for each language, we construct highly informative prompts to guide Grok-3 in predicting coreference chains and reconstructing zero anaphora. Our approach secured 3rd place in the CRAC 2025 shared task.

1 Introduction

Coreference resolution is the task of identifying and grouping linguistic expressions in a text that refer to the same real-world entity or event, which may occur within the same sentence or be separated across multiple sentences, sometimes requiring analysis of the entire document for accurate identification. This task involves two subtasks: identifying entity mentions (including zero mentions such as pro-drops) and clustering them into groups corresponding to actual entities or events.

This paper describes our approach to the CRAC 2025 Shared Task on Multilingual Coreference Resolution, which is the fourth iteration of this ongoing challenge organized in conjunction with the CODI-CRAC 2025 Workshop at EMNLP 2025. Building on the successes of previous editions in 2022 (Žabokrtský and Ogrodniczuk, 2022), 2023 (Žabokrtský and Ogrodniczuk, 2023), and 2024 (Novák et al., 2024), the 2025 shared task emphasizes multilingual capabilities and introduces a dedicated LLM Track to explore the potential of large language models (LLMs) in handling coreference across typologically diverse languages. Participants are tasked with developing systems that not only detect mentions, including the reconstruction of zero mentions but also accurately cluster them, while accommodating linguistic variations such as different annotation styles and the presence of pro-drops.

The data for the shared task is based on the public edition of CorefUD 1.3 (Novák et al., 2025), comprises 22 different datasets across 17 languages in a harmonized scheme. Compared to CRAC 2024, two additional languages, Korean and Hindi, with a new French dataset, while excluding the English-ParCorFull and German-ParCorFull datasets. The data is provided in CoNLL-U format, with coreference annotations in the MISC column, and a plaintext variant is available for LLM-based approaches to facilitate prompt engineering and in-context learning. To promote realism, development and test sets are reduced to mini-dev and mini-test splits (approximately 25,000 words each), and morpho-syntactic features in input data are generated using UDPipe 2 (Straka, 2018), simulating scenarios without gold annotations.

A key innovation in CRAC 2025 is the bifurcation into two tracks: the LLM Track, which restricts systems to primarily LLM-driven methods such as fine-tuning, in-context learning, prompt tuning, and constrained decoding, and the Unconstrained Track, which allows hybrid or non-LLM approaches. Our participation in the LLM Track leverages to address the challenges of multilingual coreference, including zero mention reconstruction and cross-lingual transfer. The evaluation employs the CorefUD scorer[1], with the primary metric being the macro-averaged CoNLL F_1 score across all datasets, using head-matching for mentions and excluding singletons. This setup encourages the development of robust, multilingual systems capable

[1] https://github.com/ufal/corefud-scorer

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 149–153
Suzhou, China, November 9, 2025.

of handling diverse linguistic phenomena.

2 Related Work

The field of coreference resolution has evolved from early rule-based and statistical models (Snyder et al., 2009) to end-to-end neural architectures that framed the task as a span-ranking problem (Wang et al., 2017). The advent of pre-trained language models like BERT (Devlin et al., 2019) further advanced the state-of-the-art, with models like SpanBERT (Joshi et al., 2020) achieving new performance benchmarks by capturing richer contextual information. However, these models still largely rely on a fine-tuning paradigm, which requires substantial annotated data, a resource scarce for most languages.

More recently, the landscape has shifted with the emergence of LLMs such as GPT-3 (Brown et al., 2020), which excel at zero-shot and few-shot learning. These models can perform complex NLP tasks through in-context learning (ICL) (Dong et al., 2024) without parameter updates, often guided by a few examples in a prompt (Chen et al., 2023). The effectiveness of ICL, however, is highly sensitive to the quality and relevance of the selected exemplars (Nie et al., 2022). While most work has relied on random or heuristic-based sample selection, our approach focuses on a difficulty-aware strategy to curate the most informative examples.

3 Method

Our approach for the LLM Track employs in-context learning through few-shot prompting and carefully designed instructions to enable LLMs to perform multilingual coreference resolution. To construct effective few-shot demonstrations, we curate samples from the mini-dev sets, which contain both raw text inputs and corresponding gold-standard annotations, making them ideal for instructional purposes.

The selection of exemplars is guided by a custom difficulty metric designed to identify challenging instances that reflect diverse linguistic phenomena, such as nominal ambiguity, pronominal interference, and zero mentions (e.g., pro-drops). This approach ensures that few-shot examples expose the LLM to complex scenarios, enhancing its robustness across the 17 languages in the dataset and potential unseen languages in the mini-test set. The difficulty score is computed as a weighted linear combination of three components: the Nominal Dissimilarity Score, the Pronoun Ambiguity Score, and the Zero Mention Score. Each component is detailed below, followed by its integration methodology.

3.1 Nominal Dissimilarity Score

The Nominal Dissimilarity Score quantifies semantic dissimilarity within coreference clusters, reflecting the resolution challenge posed by diverse or partial matches between mentions. For each text sample (in plaintext format with annotations such as [eX mentionleX]), we employ the Gemini Flash 2.0 API to execute a multi-step analysis as follows:

- **Mention Extraction**: A regular expression-based parser extracts all mentions and their corresponding cluster assignments from the annotated plaintext, producing structured JSON output containing mention spans and entity identifiers.
- **Representative Phrase Selection**: Using the extracted mentions, the LLM identifies a representative phrase for each cluster (e.g., the most descriptive or head noun phrase).
- **Semantic Similarity Computation**: For each cluster, we use Gemini Flash 2.0 with a tailored instruction prompt to compute a semantic similarity score (on a 0-100 scale) between each mention and its representative phrase, using LLM embeddings. The overall Nominal Dissimilarity Score is the average of inverted similarity scores (100 - similarity) across all mentions in the sample, reflecting resolution difficulty.

This score identifies instances where mentions within a cluster exhibit low semantic similarity, increasing resolution difficulty. In cases of processing errors (e.g., invalid LLM output), the score defaults to 0.0 in case of processing errors.

3.2 Pronoun Ambiguity Score

The Pronoun Ambiguity Score evaluates pronominal ambiguity by measuring the level of interference from distracting antecedents for pronouns within the text. We utilize the Gemini Flash 2.0 API for this analysis:

Pronoun Extraction The LLM extracts all pronouns, recording their positions and contexts, and outputs them in structured JSON format.

Relationship Analysis For each pronoun, the LLM identifies potential antecedents within a ±150 character window, categorizing them as either supporting" (those that align with the correct coreferential entity) or distracting" (plausible but incorrect alternatives based on gender, number, or semantic agreement). The ambiguity score for each pronoun is calculated as: distracting count minus supporting count (higher values indicate greater ambiguity).

Aggregation The overall Pronoun Ambiguity Score is the average of positive per-pronoun scores (where distractors outnumber supporters), normalized to a 0–100 scale by dividing by the maximum observed score in the mini-dev set and multiplying by 100. This focuses on genuinely ambiguous cases. Summary statistics, including total pronoun count and score distribution, are computed for validation purposes. In cases of processing errors (e.g., invalid LLM output), the score defaults to 0.0. This metric captures discourse-level challenges where multiple candidate antecedents can mislead the resolution process.

3.3 Zero Mention Score

The Zero Mention Score quantifies the difficulty posed by implicit references (e.g., pro-drops), which require significant syntactic and semantic inference for reconstruction. These are marked by "##" in the plaintext format. Instead of using a tiered system with arbitrary boundaries, we employ a continuous function to provide a more robust and principled score. The score is calculated as a capped linear function of the number of zero-mention occurrences (N_{zero}) within the sample:

$$S_{\text{zero}} = \min(N_{\text{zero}} \times C, 100)$$

where C is a scaling factor chosen to make the score's magnitude comparable to the other two metrics. For our experiments, we set $C = 2$.

3.4 Difficulty Score Integration

The final difficulty score (S_{diff}) is computed as a weighted linear combination of the three component scores:

$$S_{\text{diff}} = 0.4 \cdot S_{\text{nom}} + 0.4 \cdot S_{\text{pron}} + 0.2 \cdot S_{\text{zero}}$$

where S_{nom}, S_{pron}, and S_{zero} represent the Nominal Dissimilarity, Pronoun Ambiguity, and Zero Mention scores, respectively. The final score is capped at 100. The weights for this combination were established to reflect the distinct nature of the challenges that each metric captures.

Nominal Dissimilarity (S_{nom}) and Pronoun Ambiguity (S_{pron}) were assigned equal, high weights of 0.4. We posit that these metrics are the primary indicators of deep inferential complexity. S_{nom} reflects semantic challenges requiring world knowledge, while S_{pron} captures structural ambiguity at the discourse level. Prioritizing these equally ensures that we select for samples rich in complex reasoning tasks, which are most beneficial for challenging the LLM in a few-shot setting.

The Zero Mention score (S_{zero}) was assigned a lower weight of 0.2. While identifying zero anaphora is crucial, this metric primarily quantifies the *frequency* of the phenomenon rather than the reasoning complexity of a single instance. Therefore, it serves as an important secondary factor that modulates the final score but is weighted less than the core semantic and discourse challenges. This weighting scheme is deliberately designed to favour exemplars that are structurally and semantically complex, aiming to maximize the learning signal provided to the LLM.

3.5 Sample Selection and Final Inference

To accommodate linguistic diversity, we compute difficulty scores and perform sample selection on a per-language basis. For each language, we rank its mini-dev samples by their difficulty scores and select the top 3 most challenging examples as few-shot demonstrations. For unseen languages, we employ a zero-shot prompt.

The prompts are tailored to each language. Finally, the constructed prompt, containing instructions and the selected few-shot examples, is fed to **Grok-3**. Grok-3 processes the input test data to generate coreference predictions, which are subsequently converted back to CoNLL-U format using the provided text2text-coref [2] tool for evaluation.

4 Results and Discussion

Our system secured the third position among four submissions in the LLM Track, with an average CoNLL F1-score of 60.09 across all datasets. This result places our system 2.87 points behind the top-performing entry's score of 62.96. Detailed per-dataset results are presented in Table 1 in the Appendix.

[2] https://github.com/ondfa/text2text-coref

Table 1: Results and rankings of our method on the datasets in the competition.

Dataset	F1-score	Dataset	F1-score	Dataset	F1-score	Dataset	F1-score
ca_ancora	68.01 (3)	pl_pcc	66.55 (3)	no_bokmaalnarc	65.18 (3)	hbo_ptnk	45.31 (2)
cs_pcedt	56.94 (3)	es_ancora	70.52 (3)	no_nynorsknarc	63.00 (3)	fr_ancor	63.77 (3)
cs_pdt	62.96 (3)	fr_democrat	61.54 (1)	tr_itcc	56.06 (2)	hi_hdtb	66.85 (3)
de_potsdamcc	57.41 (3)	hu_szegedkoref	61.61 (2)	cu_proiel	43.74 (3)	ko_ecmt	50.32 (4)
en_gum	61.71 (3)	ru_rucor	67.59 (4)	en_litbank	69.12 (3)		
lt_lcc	65.35 (1)	hu_korkor	50.58 (2)	grc_proiel	47.86 (3)		

The strongest performance of our system was observed on the `fr_democrat` and `lt_lcc` datasets, where it achieved the top rank. We attribute this success to our use of in-context learning with preselected examples, which proved particularly effective for identifying essential mentions. Similarly, the system demonstrated competitive performance on datasets characterized by a high frequency of zero mentions—such as `tr_itcc`, `ca_ancora`, and `hu_szegedkoref` securing ranks from 2 to 3. The high CoNLL F1 scores on these datasets suggest that our few-shot approach successfully captured zero-mentions, a key objective of the shared task.

Limitations and Ablation Analysis Despite these successes, our approach has several limitations that warrant discussion. A key limitation in our post-hoc analysis is the absence of a direct baseline comparing our difficulty-aware selection against a random sampling strategy using the same Grok-3 model. Such a comparison would have precisely quantified the performance gain attributable solely to our selection methodology. While time and resource constraints of the shared task prevented this ablation study, we acknowledge its importance for a more thorough evaluation. This lack of a direct baseline is a primary area for future work.

Beyond the need for a baseline, the overall performance was constrained by several factors. Firstly, the model struggled with long input contexts, particularly evident in datasets like `fr_ancor`. When faced with extensive texts, the model often failed to maintain context over long distances, leading to the fragmentation of coreference chains and an inability to resolve long-distance dependencies. This issue was compounded by occasional failures to adhere to the specified output format, which caused critical errors during the `text2conllu` conversion phase and prevented the establishment of semantic links between distant mentions.

Secondly, our strategy's reliance on only three few-shot examples per language, while computationally efficient, likely provided the LLM with an insufficient representation of linguistic diversity. Although our metrics aimed to select the *most informative* examples, this low quantity may have constrained the model's ability to generalize across the full spectrum of coreference phenomena present in the test data.

Finally, performance was impacted by domain and language mismatches. The `ko_ecmt` dataset, with its admixture of Korean, English, and Chinese text, posed a significant challenge for our clustering algorithm. Similarly, the presence of ancient languages (`hbo_ptnk`, `cu_proiel`, `grc_proiel`), which are out-of-domain relative to the LLM's pretraining data, highlighted the difficulty of adapting modern LLMs to historically distant linguistic contexts, even with targeted few-shot prompting.

5 Conclusion

In this paper, we introduced a few-shot coreference resolution system for the CRAC 2025 Shared Task using difficulty-aware in-context learning, achieving third place in the LLM track. Our approach demonstrated the viability of no-fine-tuning methods but was limited by using only three training exemplars per language. This led to coreference chain fragmentation and reduced performance on long documents and out-of-domain ancient languages. Future work will focus on two key areas: first, establishing a clear baseline against random sampling to rigorously validate the impact of our difficulty metrics; and second, exploring methods to incorporate a larger, yet still curated, set of diverse training examples to improve generalization and overall performance.

Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuanjing Huang. 2023. KNSE: A knowledge-aware natural language inference framework for dialogue symptom status recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10278–10286.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *Preprint*, arXiv:2212.02216.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Peter Bourgonje, Silvie Cinková, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, and 23 others. 2025. Coreference in universal dependencies 1.3 (CorefUD 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 73–81.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207. Association for Computational Linguistics.

Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2017. Affinity-preserving random walk for multi-document summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 210–220.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2022. *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2023. *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*.