# Closing the Gap: Robust Multilingual Coreference Resolution with DAgger

**Thomas Morton**
University of California, San Diego
thmorton@ucsd.edu

**Alex Warstadt**
University of California, San Diego
awarstadt@ucsd.edu

## Abstract

We present DAggerCoref, our submission to the CRAC 2026 Shared Task on Multilingual Coreference Resolution. DAggerCoref is a three-stage cascade built on XLM-RoBERTa-large: a gap classifier for zero pronoun detection, a mention head classifier, and a coarse-to-fine antecedent scorer. Our central contribution is applying DAgger (Ross et al., 2011) to coreference resolution: after training the antecedent scorer on gold mentions, we fine-tune on a 50/50 mix of gold and pipeline-predicted mentions, closing the train/test distribution mismatch and improving development set macro CoNLL F1 by 1.10 points. We also introduce Otsu adaptive thresholding for zero pronoun detection, which matches gold-tuned per-dataset thresholds without requiring any gold supervision. Our system achieves a macro CoNLL F1 of 67.56 on the official test set across 27 datasets and 19 languages.[1]

## 1 Introduction

The CRAC 2026 Shared Task on Multilingual Coreference Resolution (Novák et al., 2026) evaluates systems on 27 datasets spanning 19 languages in the CorefUD format (Nedoluzhko et al., 2022). Systems must identify mentions—including reconstructing zero pronouns in pro-drop languages—and cluster them into coreference chains, scored by macro-averaged CoNLL F1 (the average of MUC, $B^3$, and $CEAF_e$).

Pipeline coreference systems typically train the antecedent scorer on gold mentions but evaluate on predicted mentions from an upstream mention detector. This creates a systematic *exposure bias*: the scorer never encounters the false positives and false negatives it will face at test time. Following Clark and Manning (2015), who first applied DAgger (Ross et al., 2011) to coreference via model stacking, we address this mismatch by fine-tuning the antecedent scorer on a mixture of gold and pipeline-predicted mentions. Additionally, we apply Otsu's method (Otsu, 1979) with a bimodality test for adaptive zero pronoun thresholding, eliminating the need for per-dataset gold-tuned thresholds.

Our contributions are:

- We apply DAgger to multilingual coreference resolution, improving macro CoNLL F1 by 1.10 over a strong focal-loss baseline by closing the gold/predicted mention gap.
- We introduce Otsu adaptive thresholding for zero pronoun detection, which matches gold-tuned per-dataset thresholds without requiring gold supervision—critical for evaluation on unseen data.
- We present a three-stage cascade with shared XLM-RoBERTa-large encoders, achieving 67.56 macro CoNLL F1 across 27 datasets and 19 languages.

## 2 Related Work

Lee et al. (2017) introduced end-to-end neural coreference resolution with marginal log-likelihood over gold antecedents, extended by Lee et al. (2018) with coarse-to-fine pruning. Dobrovolskii (2021) introduced word-level coreference, representing each mention by a single word and reconstructing spans post hoc, which we adopt. CorPipe (Straka, 2025) has won the CRAC shared task every year since 2022.

Clark and Manning (2015) first applied DAgger (Ross et al., 2011) to coreference resolution, training a cluster-merging policy for an entity-centric agglomerative clusterer. We extend this approach to the multilingual setting with a modern XLM-RoBERTa encoder (Conneau et al., 2020), applying DAgger at the antecedent scoring stage rather than at cluster-merging.

[1] Code is available at https://github.com/tgmorton/dagger-coref.



*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 217–221

## 3 System Description

DAggerCoref is a three-stage cascade where each stage is an XLM-RoBERTa-large classifier (Conneau et al., 2020). Stages 2 and 3 warm-start their encoders from the trained Stage 1 gap classifier, sharing a common multilingual representation. Figure 1 shows the pipeline.

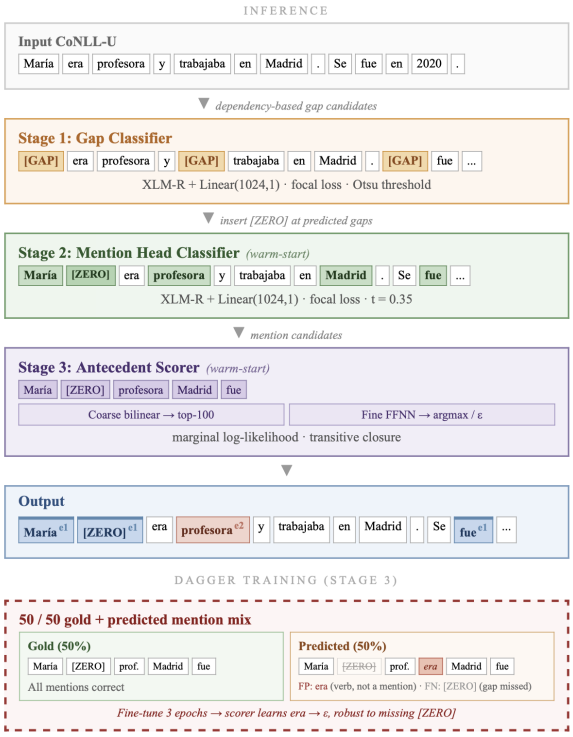


Figure 1: DAggerCoref pipeline. Stages 2 and 3 warm-start from Stage 1. Bottom: DAgger training exposes the scorer to predicted-mention noise (false positives and false negatives).

### 3.1 Stage 1: Gap Classifier

We detect zero pronouns by inserting learned `[GAP_i]` special tokens at syntactically plausible positions identified from the UD dependency parse (verbs missing overt subjects, transitive verbs missing objects, nouns missing possessors). The encoder's hidden states at `[GAP_i]` positions are passed through a linear classifier to predict whether each gap is a real zero pronoun. We train with focal loss (Lin et al., 2017) ($\gamma = 2$, $\alpha = 0.25$) to handle the class imbalance ($\sim$1:14 positive-to-negative ratio), achieving 94.1% average F1 on zero-rich datasets.

**Otsu adaptive thresholding.** At inference, we determine the classification threshold adaptively using Otsu's method (Otsu, 1979), which finds the threshold that maximizes the between-class variance of the gap probability distribution—effectively separating a cluster of high-probability zeros from low-probability non-zeros. Before applying Otsu's method, we test whether the distribution is genuinely bimodal by comparing 1- vs. 2-component Gaussian mixture models via BIC, requiring both a significant fit improvement and component separation of $>3\sigma$. Languages with genuine zeros produce bimodal distributions; non-pro-drop languages produce unimodal distributions near zero, for which the bimodality test returns a conservative fallback of $t = 0.5$ (the sigmoid midpoint), yielding no predicted zeros. This eliminates per-dataset gold-tuned thresholds entirely.

### 3.2 Stage 2: Mention Head Classifier

A token-level binary classifier predicts whether each word is a coreference mention head, replacing rule-based POS filtering. The encoder is warm-started from Stage 1. We train with focal loss and apply a fixed global threshold of $t = 0.35$ (tuned on dev). A per-dataset sweep found this value optimal across all 27 datasets, leaving no signal for adaptive thresholding to exploit. The MHC threshold also does not optimize mention F1 in isolation: the scorer can filter false-positive mentions via its epsilon score, but false negatives are irrecoverable, biasing the optimal threshold below any data-driven class boundary that Otsu would identify. All predicted `[ZERO]` tokens from Stage 1 are automatically included as mention candidates.

### 3.3 Stage 3: Antecedent Scorer

We use a coarse-to-fine antecedent ranking architecture (Lee et al., 2018) with head-only mention representations (Dobrovolskii, 2021). For each mention $i$, a bilinear scorer $g_i^\top W g_j$ prunes to the top-$k$ antecedent candidates, which are then scored by a two-layer FFNN on concatenated features $[g_i; g_j; g_i \odot g_j]$. A learned epsilon scorer predicts per-mention "new entity" scores. We train with marginal log-likelihood over all gold antecedents (Lee et al., 2017).

Documents are segmented into 512-subword windows (384 content + 128 overlap), with segment length informed by Joshi et al. (2019). At inference, we replace the fixed sliding grid with mention-centered windows that place each focus mention near the window's end, devoting $\sim$460 of 512 subwords to backward context where an-

tecedents reside. We also widen top-$k$ from 50 (training) to 100 (inference) to reduce entity fragmentation. Cross-segment coreference is resolved via transitive closure over pairwise antecedent links.

### 3.4 DAgger Training

Standard antecedent scorers are trained on gold mentions but evaluated on noisy predicted mentions from upstream stages. In our system, the gold-mention upper bound is 77.9 macro CoNLL F1 on the development set, while the best end-to-end pipeline before DAgger scores 66.77—a gap of 11.1 points attributable to this exposure bias.

We apply a single round of DAgger-style data aggregation (Ross et al., 2011; Clark and Manning, 2015) to close this gap:

1. Train the baseline antecedent scorer on gold mentions (10 epochs).
2. Run the full pipeline (Stages 1 + 2) on the training data to produce predicted mentions with realistic noise.
3. Merge the resulting predicted-mention JSONL files with the original gold-mention JSONL at a 50/50 document ratio, with document-level train/dev splitting to prevent leakage.
4. Fine-tune the baseline scorer on this mixed dataset for 3 epochs with conservative learning rates (encoder: $2\times10^{-6}$, heads: $5\times10^{-5}$) and a brief encoder freeze (200 steps).

The DAgger-trained scorer encounters false-positive mentions (predicted but not gold) during training, learning to assign them the epsilon (new entity) antecedent rather than erroneously linking them. This contributes +1.10 macro CoNLL F1 over the focal-loss baseline. Since the 50/50 mix was chosen naively without sweeping the ratio or exploring multi-round schedules, +1.10 likely represents a floor on the benefit DAgger can provide for this task; determining the optimal data curriculum is left to future work.

## 4 Training

All stages use XLM-RoBERTa-large (560M parameters) with bf16 mixed precision on a single NVIDIA A100 GPU. Stage 1 trains for 10 epochs (∼30 min, effective batch 32, lr $2\times10^{-5}$). Stage 2 trains for 5 epochs (∼60 min, effective batch 32, focal loss). Stage 3 trains for 10 epochs (∼4 hours, effective batch 4) with a two-stage curriculum: the encoder is frozen for 500 steps while the scoring heads warm up, then unfrozen with reduced lr ($1\times10^{-5}$). DAgger fine-tuning adds 3 epochs (∼2 hours). Total training time is approximately 7 hours on a single A100. No ensembling is used in the final submission.

| | System | Macro CoNLL F1 |
|---|---|---|
| *Unconstrained track* | | |
| 1 | CorPipe-ensemble | 77.11 |
| 2 | CorPipe-single | 76.18 |
| 3 | CorPipe-single-lg | 72.32 |
| 4 | **DAggerCoref (ours)** | **67.56** |
| 5 | Stanza-coref | 67.00 |
| 6 | Baseline | 54.54 |
| 7 | AU-KBC | 35.24 |
| *LLM track* | | |
| 1 | Antoine Bourgois | 74.32 |
| 2 | hejmanj | 73.83 |
| 3 | portnlp | 68.69 |
| 4 | pavlk-mm | 46.19 |

Table 1: Official test set results (macro-averaged CoNLL F1 across 27 datasets).

## 5 Results

DAggerCoref achieves a macro CoNLL F1 of 67.56 on the official test set, ranking 4th in the unconstrained track (Table 1; full leaderboard in Novák et al., 2026). We outperform the next unconstrained-track entry by +0.56 points and the shared task baseline by +13.02 points.

Table 2 shows our per-dataset breakdown on the test set. Our system is strongest on Romance languages and Russian (ru_rucor 79.12, es_ancora 77.39, ca_ancora 76.77) and weakest on Turkish (tr_itcc 42.92), ancient languages (hbo_ptnk 57.90, cu_proiel 57.20), and la_coreflat (55.95).

The gold-mention upper bound for our antecedent scorer is 77.9 macro CoNLL F1 on development data. The remaining 10.3-point gap to the end-to-end test score (67.56) is due to mention detection errors, indicating that improving mention identification remains the primary bottleneck.

## 6 Analysis

### 6.1 Component Contributions

Table 3 shows the cumulative contribution of each inference and training decision on the development set. Starting from a fixed-grid sliding win-

| Dataset | F1 | Dataset | F1 |
|---|---|---|---|
| ru_rucor | 79.1 | cs_pdtsc | 68.2 |
| es_ancora | 77.4 | en_gum | 68.6 |
| ca_ancora | 76.8 | fr_litbankfr | 66.9 |
| cs_pdt | 74.6 | lt_lcc | 66.5 |
| pl_pcc | 73.7 | grc_proiel | 65.2 |
| en_litbank | 73.6 | nl_openboek | 64.5 |
| en_fantasy | 73.0 | ko_ecmt | 61.1 |
| no_bokmaal | 72.9 | hu_szegedkoref | 61.0 |
| hi_hdtb | 72.8 | hu_korkor | 60.5 |
| no_nynorsk | 72.0 | hbo_ptnk | 57.9 |
| de_potsdamcc | 71.9 | cu_proiel | 57.2 |
| cs_pcedt | 70.8 | la_coreflat | 56.0 |
| fr_ancor | 70.2 | tr_itcc | 42.9 |
| fr_democrat | 69.0 | | |
| **Average** | | | **67.56** |

Table 2: Per-dataset macro CoNLL F1 on the test set, sorted by score.

| Configuration | Dev F1 | Δ |
|---|---|---|
| Gold mentions (ceiling) | 77.90 | — |
| Fixed-grid windows, $k=50$ | 64.09 | — |
| + Centered windows | 66.17 | +2.08 |
| + Antecedent widening ($k=100$) | 66.77 | +0.60 |
| + DAgger fine-tuning | **67.87** | +1.10 |

Table 3: Cumulative ablation on the development set. Each row adds one component to the row above it.

dow baseline (64.09), mention-centered windows contribute +2.08 by improving antecedent visibility; widening the top-$k$ from 50 to 100 adds +0.60 by reducing entity fragmentation on long documents (gains plateau at $k = 100$; $k = 150$ yields identical F1); and DAgger fine-tuning adds a further +1.10 by exposing the scorer to realistic pipeline noise during training. The gold-mention upper bound (77.9) indicates that ∼10 points remain lost to mention detection errors.

### 6.2 Cross-Lingual Otsu Thresholding

To validate that Otsu thresholding generalizes to unseen languages, we trained gap classifiers with individual languages held out and evaluated zero detection using only the Otsu-determined threshold. Table 4 shows the results. On held-out Spanish, Otsu predicts 391 zeros vs. 393 gold (delta of 2); on held-out Czech, the worst-case error is 44/1770 (2.5%) on cs_pdtsc. For non-pro-drop languages (e.g., en_litbank, ko_ecmt), the bimodality test correctly returns a conservative fallback, predicting zero zeros. These results confirm that Otsu thresholding eliminates the need for per-dataset gold tuning while maintaining near-perfect zero counts even on languages the classifier has never seen.

| Held out | Dataset | Pred | Gold | F1 |
|---|---|---|---|---|
| Spanish | es_ancora | 391 | 393 | 95.2 |
| Czech | cs_pcedt | 685 | 706 | 93.6 |
| | cs_pdt | 596 | 589 | 89.3 |
| | cs_pdtsc | 1726 | 1770 | 94.1 |

Table 4: Cross-lingual zero-shot gap detection with Otsu thresholding. The classifier was trained without the held-out language. Pred/Gold are zero counts; F1 is gap classification F1.

| Antecedent distance | Accuracy |
|---|---|
| Same sentence | 92.4% |
| 1 sentence | 88.8% |
| 2–3 sentences | 81.5% |
| 4–10 sentences | 65.9% |
| 11+ sentences | 45.0% |

Table 5: Antecedent linking accuracy by distance between mention and closest gold antecedent (dev set).

### 6.3 Error Analysis

Our system's primary failure mode is *entity fragmentation*: gold entities are split into multiple predicted entities when long-distance antecedent links are missed. On the development set, the system predicts 30,575 entities vs. 15,622 gold (1.96× over-splitting), and 25.2% of gold entities are split across two or more predicted entities. Table 5 shows that linking accuracy degrades sharply with antecedent distance: same-sentence links are resolved at 92.4%, but accuracy drops to 45.0% for antecedents more than 10 sentences away. This suggests that future work should focus on long-range coreference, whether through larger context windows, entity-level representations, or iterative refinement.

## 7 Conclusion

We presented DAggerCoref, a three-stage multilingual coreference system that closes the train/test mention-distribution gap via DAgger training. Combined with Otsu adaptive thresholding for zero pronouns, it achieves 67.56 macro CoNLL F1 across 27 datasets and 19 languages. Future work includes iterative DAgger rounds and joint multi-stage training to reduce inter-stage error propagation.

## Limitations

Our system uses a single XLM-RoBERTa-large encoder (560M parameters) without ensembling, which limits performance relative to systems using larger encoders. The dependency-based gap candidate generation covers only 79% of gold empty nodes, imposing a recall ceiling on zero pronoun detection for languages with non-standard zero deprels (e.g., Czech `nmod:gen`). Finally, our DAgger procedure uses a single round; iterative rounds may yield further gains but were not explored due to time constraints.

## Acknowledgments

This work used resources available through the National Research Platform (NRP) at the University of California, San Diego. NRP has been developed, and is supported in part, by funding from National Science Foundation, from awards 1730158, 1540112, 1541349, 1826967, 2112167, 2100237, and 2120019, as well as additional funding from community partners.

## References

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

Anna Nedoluzhko, Michal Novák, Silvie Cinková, Zdeněk Žabokrtský, and 1 others. 2022. CorefUD 1.0: Coreference meet universal dependencies. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2026. Findings of the fifth shared task on multilingual coreference resolution: Expanding datasets for long-range entities. In *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)*, San Diego, California, USA. Association for Computational Linguistics.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 627–635.

Milan Straka. 2025. CorPipe at CRAC 2025: Multilingual coreference resolution and a detailed ablation study. In *Proceedings of the CODI-CRAC 2025 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics.