MSCAW-coref: Multilingual, Singleton and Conjunction-Aware Word-Level Coreference Resolution

Houjun Liu*, **John Bauer**, **Karel D'Oosterlinck**
Christopher Potts, **Christopher D. Manning**
Stanford University
*houjun@stanford.edu

Abstract

Modern multi-lingual coreference resolution approaches largely focus on the clustering of mention spans, leading to quartic complexity in the choice of both spans and span links. The recently published CAW-coref reduces coreference complexity to quadratic while still attaining 97.9% of SOTA performance through a word-level approach on the English OntoNotes slice. Naively extending the CAW-coref algorithm towards multiple languages on the CorefUD dataset results in a lackluster 77.4% of SOTA performance. We find this is due to annotation differences across OntoNotes and CorefUD—the latter features singletons which CAW-coref is not able to classify. In response, we introduce MSCAW-coref, which extends CAW-coref to work in a multilingual setting and accounts for singleton mentions. We demonstrate that MSCAW-coref attains 95.7% of SOTA performance on CorefUD while being substantially more efficient. Our algorithmic contribution towards accounting for singletons is a major driver of performance. Finally, we discuss the cross-linguistic generalization capability of our approach. We release the models, code, and a package for performing coreference analysis for the community as a part of Stanza (https://github.com/stanfordnlp/stanza).

1 Introduction

Coreference resolution ("coref") is the task of finding textual spans within a document that refer to the same entity in the real world. It is an important parsing step with many applications in NLP (Jurafsky and Martin, 2021). Coref is especially difficult when processing long documents with corresponding long chains of dependencies. Classical end-to-end neural approaches (Lee et al., 2017) often use a procedure that resolves coref by first identifying spans and then linking them together, leading to an $O(n^4)$ computation for n tokens. Worse yet, state-of-the-art (SOTA) coref approaches are often transition parsers (Bohnet et al., 2023), which require multiple forward passes of a language model (LM) to resolve all chains. Such inefficient computation is often untenable, especially in long documents.

Dobrovolskii (2021) and D'Oosterlinck et al. (2023) introduce WL-coref and CAW-coref, which are two iterations of an approach which (1) creates word-level bilinear links for head-word identification, (2) filters the links for likely coreference, and (3) extracts the spans surrounding each headword. This only-once-bilinear approach reduces the complexity of the coref computation to $O(n^2)$ while causing little loss in coref performance.

While these approaches are promising for high-efficiency coref computations, two limitations remain: first, these current approaches only focus on English, usually using the OntoNotes corpus (Weischedel et al., 2011); second, the identification of singleton mentions are beneficial across application domains of coreference (Recasens et al., 2013) but cannot be represented with existing word-level approaches due to the current heuristic of non-mentions being words with no antecedents.

In response, we introduce MSCAW-coref, an extension of the word-level coreference approach that addresses both of these challenges. To support singleton links, we revise the head-word linking step in CAW-coref to include a "sequence start" antecedent link for all first references in a chain, thereby supporting singletons through having at least one antecedent link; to support multilinguality, we apply a low-rank adaptation parameter-efficient fine-tuning scheme to XLM-RoBERTa (Hu et al., 2021; Conneau et al., 2020) to create contextual embeddings with multilingual support.

We train our approach on CorefUD, a multilingual coreference dataset with annotated singletons (Nedoluzhko et al., 2022), and demonstrate 95.7% performance compared to the best-reported quartic multilingual results while maintaining the

Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2024), pages 33–40
November 15, 2024. ©2024 Association for Computational Linguistics

dramatically more efficient modeling approach of word-level coref. We further demonstrate that our approach can zero-shot generalize to unseen languages at training time at a slight cost to performance.

2 Related Work

2.1 Modeling Approaches

Transition and Sequence-to-Sequence Parses The current state-of-the-art in coref (Bohnet et al., 2023) is formulated as an autoregressive, transition-based parser which creates each link with a forward pass of a 13B parameter LM until the reference chains are built. These methods have been demonstrated to generalize well over structured language parsing tasks (Paolini et al., 2021) and can be reformulated as autoregressive language modeling tasks either by identifying coreferences directly (Zhang et al., 2023) or through many surrogate tasks such as question-answering (Wu et al., 2020) or even language model prompting (Le and Ritter, 2023). While the performance of these approaches is strong, processing n tokens corresponds to worst-case n forward passes with a time complexity of $O(n)$ of a full (possibly very large, as in the case of Bohnet et al., 2023) LM required building all transitions, which introduces significant inefficiencies for long documents.

Span-Level Parses Despite the significant performance gains of recent Seq2Seq approaches, the vast majority of modern approaches are span-level parses which first formulate likely mentions before linking them together. The first end-to-end coreference model (Lee et al., 2017) follows this approach, which was later improved with an LM for contextual embeddings (Joshi et al., 2019) and multilingualism (Pražák et al., 2021). In addition to span-level linking, later work such as SpanBERT (Joshi et al., 2020) improved the performance even further by incorporating span-level representations. While being significantly more scalable than transition-based parses, these approaches still require the LM to disambiguate coref decisions, scaling by a factor of $O(n^4)$ for n input tokens (with pruning to optimize runtime performance at the cost of accuracy and to keep the problem from being intractable) due to the need to first create spans $O(n^2)$ then link them together $O(n^2)$.

Word-Level Parses In response to these inefficiencies, approaches emerged that link words together first prior to detecting spans. Kirstain et al. (2021) achieved promising span-level results without using spans at all, by formulating a word-level link to the end of each span instead. In this work, we build most directly upon `WL-coref` and `CAW-coref` (Dobrovolskii, 2021; D'Oosterlinck et al., 2023)—approaches that link head-words together before expanding each into spans.

2.2 Multilinguality

Recent approaches that demonstrated performance gains in handling multilinguality vary from language-specific fine-tuning (Skachkova et al., 2023), monolingual training from scratch (Pražák et al., 2021), or joint training with a multilingual LM (Straka, 2023). Despite the gains from specific fine-tuning demonstrated by prior approaches, the joint training method currently holds the best result for the multilingual coreference shared task (Žabokrtský et al., 2023) and is extended upon in this work.

3 `MSCAW-coref`

3.1 Data Preprocessing

To create head-word coreference data via annotated span-level entities, we follow `CAW-coref`. We use the dependency parse information given in the source dataset to pick the headword that is (1) dependent on a word outside the span or, if available, (2) coordinating conjunction within each span, if less than two dependency steps away from the headword from (1). We discuss concerns of soundness for maintaining conjunction-awareness across languages in appendix C.

3.2 Modeling

Our `MSCAW-coref` extends `CAW-coref` (D'Oosterlinck et al., 2023). We now describe our approach here while additionally summarizing the aspects of `CAW-coref` left unchanged.

Word-Level Representations `CAW-coref` leveraged a monolingual LM backbone, specifically `RoBERTa-large` (Liu et al., 2019), for contextual word-level representations by performing a single forward pass of the input document. To support multilingualism, we elected to use the larger 561M parameter `XLM-RoBERTa-large` (Conneau et al., 2020) as our LM backbone. To improve training time performance, we tune our approach using Low-Rank adaptations (Hu et al., 2021).

Coarse Scoring Without change from CAW-coref, a *coarse antecedent score* is created by a bilinear mapping between each of the input word embeddings obtained in the previous step. For each word, then, the top k coarse antecedents' embeddings are then passed to the next step.

Final Scoring and Singleton Prediction We first apply a small feed-forward network to compute a *fine antecedent score* for each word against its top k coarse antecedents, with higher values representing headwords that are more likely to be coreferential.

Second, we formulate an additional binary classification task whereby the fine antecedent scores of all words, including those in the future, are used as input features to predict whether or not each word is the first occurrence of a coreference chain.

After this is complete, for each word in the document, we obtain (1) k real-valued *antecedent score*s—computed as the sum of the rough and fine antecedent scores—for being a possible antecedent corresponding to k candidate antecedents in the document as well as (2) a single real-valued score for that word being the first member of a mention.

Coref Chain Construction We perform a greedy breadth-first search procedure using the scores computed in the fine-scoring step to chain corefs. We first examine the highest score for each word and delineate three cases—(1) if all of its scores are negative, we consider the word not coreferent and ignore it; (2) if any of its top-k antecedent scores are the highest of all scores, we add the corresponding antecedent word to our search stack; (3) if the first-mention score is the highest, we mark that word as the first mention in our search tree and add it to the search stack. After emptying the search stack, we obtain chains of coreferent words by retracing antecedent links, with the first token of each chain marked as "first-mention".

Notably, we can detect singletons by distinguishing cases (1) and (3)—words could have no valid antecedents (i.e., fitting case (2)), yet still, be added to our search/coref stack—even if size 1—due to its first-mention score.

Span Extraction Finally, exactly following CAW-coref, for each coreferent word, a span is extracted using a feed-forward neural network followed by a 1-dimensional convolutional layer which marks the start and end of each span. Coreference cluster information is not given to this step.

4 Experiments

4.1 Data

Most current approaches to coref are trained on OntoNotes (Weischedel et al., 2011) (including previously CAW-coref and WL-coref), which is a corpus which both does not include support for singletons and have fairly shallow coverage of both languages and linguistic phenomena (Nedoluzhko et al., 2022; Zeldes); the dataset includes only English, Arabic, and Chinese sections.

However, recent advances in universal syntactical tagging (de Marneffe et al., 2021) resulted in much more standardized annotations of morphological features as well as dependencies (necessary for our approach) across languages, leading to the development of CorefUD (Nedoluzhko et al., 2022): a multilingual corpus for coreference resolution. This corpus is suitable for training our current task as CorefUD has support for a variety of languages (10) spanning across the Germanic, Slavic, and Romance families, and has annotations for singleton mentions. Further, as described in section 4.2, the corpus has been widely used in shared tasks for multilingual coref.

To train and evaluate our model, we select the entire publically available subset of CorefUD published for the CRAC shared task, and prepare the dataset in the manner described further in section 4.2. We use train/dev splits provided by the shared task, and make no modification in terms of the data subset selection; if multiple datasets were available for a particular language, we mixed together all of them and trained jointly.

4.2 Baseline Study

Baselines The CRAC shared task on multilingual coreference resolution (Žabokrtský et al., 2023) directly uses the CorefUD (Nedoluzhko et al., 2022) dataset; approaches presented in the task, therefore, provide suitable and timely baselines for multilingual coreference resolution. We therefore elect to score our approach against the top-performing approaches presented in that shared task. We also benchmark applying CAW-coref directly with a multilingual backbone without the proposed changes for coref chain construction and singletons.

Scoring MSCAW-coref follows a *different* definition of head-words (due to conjunction resolution described in section 3.1). This makes exact

	efficiency		MUC			B^3			ceaf$_e$			mean
	complexity	LM params	R	P	F1	R	P	F1	R	P	F1	F1
ours	$O(n^2)$	561M	0.782	0.760	0.771	0.74	0.748	0.744	0.717	**0.764**	0.740	0.752
ours (naive CAW-coref)†	$O(n^2)$	561M	0.777	0.773	0.775	0.530	0.729	0.613	0.306	0.746	0.434	0.608
Straka, 2023	$O(n^4)$	1.2B	**0.810**	**0.814**	**0.812**	**0.779**	**0.780**	**0.78**	**0.788**	0.741	**0.763**	**0.785**
Anonymous‡	-	-	0.751	0.803	0.776	0.715	0.773	0.743	0.750	0.725	0.737	0.750
Pražák and Konopik, 2022*	$O(n^4)$	561M	0.728	0.762	0.745	0.658	0.639	0.649	0.637	0.523	0.574	0.656
Pražák et al., 2021	$O(n^4)$	179M	0.642	0.776	0.703	0.422	0.714	0.531	0.255	0.702	0.374	0.536

Table 1: Performance of our approach on the CorefUD 1.1 dataset against baseline and top performers from the 2023 CRAC multilingual shared task, dev slice (Nedoluzhko et al., 2022). **mean F1** is the main metric being evaluated. Scores are calculated with the official scorer of the CRAC shared task but using **exact span matches** and **including singletons**. Where possible, the published dev predictions from the shared task are used. †: implementation of `CAW-coref` with our proposed multi-lingual backbone without novel singleton scorer. ‡: anonymous submission to 2023 challenge without corresponding publication. *: results presented are an iteration included in the 2023 shared task. Model optimization details are given in appendix B.

Span LEA / Held Out		Germanic			Romance			Slavic			Uralic
	all	**no**	**de**	**en**	**es**	**fr**	**ca**	**pl**	**ru**	**cs**	**hu**
none	0.689	0.734	0.638	0.656	0.712	0.503	0.693	0.68	0.677	0.715	0.569
no	-0.075	-0.054	-0.144	-0.038	+0.043	+0.061	+0.033	+0.037	+0.001	+0.008	+0.084
de	-0.085	-0.106	-0.317	-0.072	+0.059	+0.060	-0.010	+0.033	+0.020	+0.020	+0.067
en	-0.074	-0.086	-0.146	-0.148	+0.088	+0.084	-0.003	+0.026	+0.008	+0.041	+0.058
es	-0.092	-0.080	-0.100	-0.062	-0.008	+0.043	+0.022	+0.032	+0.024	-0.007	+0.005
fr	-0.163	-0.052	-0.106	-0.054	+0.050	-0.098	+0.001	+0.012	+0.031	+0.017	+0.042
ca	-0.076	-0.081	-0.119	-0.025	-0.007	+0.067	-0.066	-0.001	+0.022	+0.039	+0.035
pl	-0.091	-0.084	-0.073	-0.049	+0.034	+0.056	+0.046	-0.307	-0.009	+0.012	+0.042
ru	-0.097	-0.073	-0.106	-0.046	+0.043	+0.063	-0.011	-0.008	-0.312	+0.025	+0.089
cs	-0.100	-0.095	-0.037	-0.029	+0.046	+0.058	+0.049	+0.022	+0.039	-0.467	+0.061
hu	-0.086	-0.095	-0.092	-0.027	+0.075	+0.049	-0.015	-0.012	+0.024	+0.017	-0.136

Table 2: Ablation of performance of `MSCAW-coref` across languages and when generalizing to unseen languages. The top row of the table shows percentage performance in span-match LEA (Moosavi and Strube, 2016); the colored rows show the percentage change in performance when the language outlined in the row is withheld from training. Results reported balanced per language. Model optimization details are given in appendix B.

head-word match (used originally in the shared task) an unsuitable metric for scoring the results obtained here; furthermore, the comparison score in the shared task does not account for singletons, which have important and distinct uses in discourse (De Marneffe et al., 2015) from regular mentions. As such, our baseline scores against CorefUD use the *exact span level matches* which also *includes singletons* instead of the head-word-only and non-singleton scores used as the primary metric of the CRAC shared task.

Notably, there is an exact algorithmic solution provided by the shared task[1] to derive the head-word from the dependency tree, so the exact span resolution task (unlike previously the partial span resolution task) is a superset of the metric usually given in the shared task.

Scores are computed with the official scoring system given in the shared task, and recomputed from published dev set outputs of shared task participants when needed.

4.3 Ablation Study

We also evaluate the performance of our model across languages and its ability to generalize to unseen languages. To do this, we sample a 10% test split from the train split of CorefUD, controlling for an equivalent representation of each language across all datasets. Then, we withhold one language at a time during training and report evaluation results across all languages (including the withheld language).

5 Results

Table 1 gives the results of our baseline study. While our approach achieves 96% of the performance of the leading solution of the shared task (Straka, 2023) on the CoNLL-2012 metric evaluated with singletons and exact span matches, we did so with significantly reduced computational complexity from $O(n^4)$ to $O(n^2)$ as well as lowered constant-time performance due to the reduc-

[1] `https://github.com/udapi/udapi-python/blob/master/udapi/block/corefud/movehead.py`

tion of parameters in the LM backbone. Notably, the highest-performing approach in the shared task using our same LM backbone (Pražák and Konopik, 2022) achieved a dramatically lower performance of 65.6% compared to our 75.2%. Furthermore, naively applying the original CAW-coref using a multi-lingual backbone, on the other hand, only results in 60.8% mean F1 compared to our 75.2% mean F1 (row 2).

We further investigate the language-specific and out-of-domain generalization results of our scheme in table 2. Results appear to be roughly clustered by language family. Romance languages generalize well amongst each other: holding out French entirely during training but including Spanish and Catalan only results in a 9.8% reduction in French performance, and holding out Spanish or Catalan at training only results in less than 1% reduction in the test performance of the other; Germanic languages appears to benefit from inclusion of all data; and Slavic and Uralic languages benefited from the *removal* of other families' languages during training. We find the performance degradation between language family lines qualitatively supported by previous work (Pražák et al., 2021)—in part due to differing annotation standards (Porada et al., 2024)—and also underscore our approach's ability to generalize zero-shot to unseen languages.

6 Conclusion

In this work, we extend CAW-coref (D'Oosterlinck et al., 2023), an instance of WL-coref (Dobrovolskii, 2021), to add support for singleton mentions and non-English languages. We did so by introducing MS-CAW coref, a modeling approach that retains word-level time-complexity while achieving performance that is within 5% of the best-performing multilingual model on the CorefUD multilingual dataset in span-match metrics. We further release our trained multilingual models and corresponding source code for use by the wider community.

Limitations

Our approach predicts singletons through disambiguation of the starts of mention chains, yet prior work (De Marneffe et al., 2015) discussed the reduction of modeling complexity through predicting coreferent sequences and singletons as separate objects. Early empirical results (appendix A) indicate that our approach performs slightly better compared to using the cluster start classifier to predict singletons only; yet, further investigations into these results would add to the understanding of coreference modeling.

Furthermore, we inherit the choice from CAW-coref that each span can be isomorphically mapped to a headword—this is not true: there will always be more spans than headwords in a sequence. Further investigations into the deduplication of overlapping spans will likely bring further gains in performance to our approach.

Recent work highlights that differing annotation standards between datasets may contribute to variations in performance in coreference tasks (Porada et al., 2024). Correspondingly, we did observe generalization differences across datasets. A systemic error analysis that takes into account these different standards can help improve the generalization performance of the approach.

Lastly, as discussed in appendix C, we note that the conjunction-awareness properties of CAW-coref did not result in performance gains of similar magnitude in the multilingual setting. Further work can investigate language-specific properties of CAW and adapt the approach for further performance improvements.

Acknowledgments

We thank our anonymous reviewers for their valuable and gracious feedback on our work. Our model is implemented for and with the tooling provided by Stanza (Qi et al., 2020).

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

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.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Marie-Catherine De Marneffe, Marta Recasens, and Christopher Potts. 2015. Modeling the lifespan of discourse entities with application to coreference resolution. *Journal of Artificial Intelligence Research*, 52:445–475.

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.

Karel D'Oosterlinck, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder. 2023. CAW-coref: Conjunction-aware word-level coreference resolution. In *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*, pages 8–14, Singapore. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

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

Daniel Jurafsky and James H Martin. 2021. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 14–19, Online. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *Preprint*, arxiv:2305.14489 [cs].

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

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.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Ma Jie, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto, et al. 2021. Structured prediction as translation between augmented natural languages. In *ICLR 2021-9th International Conference on Learning Representations*, pages 1–26. International Conference on Learning Representations, ICLR.

Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Cheung. 2024. Challenges to evaluating the generalization of coreference resolution models: A measurement modeling perspective. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15380–15395, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Ondřej Pražák and Miloslav Konopik. 2022. End-to-end multilingual coreference resolution with mention head prediction. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 23–27, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ondřej Pražák, Miloslav Konopík, and Jakub Sido. 2021. Multilingual coreference resolution with harmonized annotations. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1119–1123, Held Online. INCOMA Ltd.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In

Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 627–633, Atlanta, Georgia. Association for Computational Linguistics.

Natalia Skachkova, Tatiana Anikina, and Anna Mokhova. 2023. Multilingual coreference resolution: Adapt and generate. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 19–33, Singapore. Association for Computational Linguistics.

Milan Straka. 2023. ÚFAL CorPipe at CRAC 2023: Larger context improves multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 41–51, Singapore. Association for Computational Linguistics.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Amir Zeldes. Opinion piece: Can we fix the scope for coreference?: Problems and solutions for benchmarks beyond OntoNotes. 13(1):41–62.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.

A Singletons vs. Starts of Sequences

Table 3 highlights that our approach performs slightly worse when using the cluster-start classification scheme discussed in section 3.2 to learn starts of sequences and singletons separately. Note that, while our strong performance is maintained in both approaches, predicting singletons resulted in a slight decrease in dev set accuracy.

B Implementation

We train all reported instances of our model using Huggingface's implementation of `xlm-roberta-large` (Wolf et al., 2020), leaving $k = 50$ rough antecedents before fine scoring. To improve training time efficiency, we restrict trainable parameters in the LM backbone using LoRA ($r = 32, \alpha = 16$) (Hu et al., 2021). The rest of the model is tuned fully. We chose a reduced learning rate for our LM backbone at $LR = 2.5 \times 10^{-5}$ with our parsing head being tuned at $LR = 3 \times 10^{-4}$.

C Scaling Conjuction Awareness to a Multilingual Setting

The conjuction-aware data preparation scheme, described in section 3.1, was originally designed with the OntoNotes English dataset (Weischedel et al., 2011). Therefore, it is apt to investigate whether the dependency-based head-word revision scheme is appropriate as the model is scaled across new languages.

Table 4 highlights that the CAW scheme empirically creates minimal (but non-zero) improvements in span-level LEA. We elected to preserve this method across all languages as a word-level approach without CAW would be unable to simultaneously resolve conjoined mentions and their constituent parts such as "Tom and Mary" simultaneously with "Tom" and "Mary" (D'Oosterlinck et al., 2023)—a condition made more frequent by the awareness of singleton mentions in the dataset.

	MUC			**B^3**			**ceaf$_e$**			**mean**
	R	P	F1	R	P	F1	R	P	F1	F1
ours	**0.782**	0.76	**0.771**	**0.74**	0.748	**0.744**	0.717	**0.764**	0.74	**0.752**
ours (singletons seperate)	0.78	0.76	0.77	0.739	0.748	0.743	**0.722**	0.758	0.74	0.751

Table 3: Performance of our approach on the CorefUD 1.1 dataset against our approach but while predicting singletons separately from mention chain starts, dev slice (Nedoluzhko et al., 2022). **mean F1** is the main metric being evaluated. Scores are calculated with the shared task scorer using **exact span matches** and **including singletons**.

	Span-Level LEA
ours	**0.689**
ours (non-CAW)	0.681

Table 4: Performance of our conjunction-aware approach on the CorefUD 1.1 dataset against our approach but while using CorefUD gold head-words.