McGill at CRAC 2023: Multilingual Generalization of Entity-Ranking Coreference Resolution Models

Ian Porada and **Jackie Chi Kit Cheung**

Mila, McGill University
{ian.porada@mail, jcheung@cs}.mcgill.ca

Abstract

Our submission to the CRAC 2023 shared task, described herein, is an adapted entity-ranking model jointly trained on all 17 datasets spanning 12 languages. Our model outperforms the shared task baselines by a difference in F1 score of +8.47, achieving an ultimate F1 score of 65.43 and fourth place in the shared task. We explore design decisions related to data preprocessing, the pretrained encoder, and data mixing.

1 Introduction

The goal of the CRAC 2023 shared task (Žabokrtský et al., 2023) is to evaluate coreference resolution models on the CorefUD 1.1 collection of datasets (Novák et al., 2022). In this paper, we describe our submission to the task, which is an adaptation of the entity-ranking model described in Toshniwal et al. (2020) with some exploration of the design decisions needed to apply this model to multiple datasets spanning multiple languages. Our final submission achieves fourth place out of nine submissions based on head-match F1 score, and third place based on exact-match F1 score.

The CRAC 2023 shared task is specifically based on the public portion of CorefUD 1.1, which includes 17 datasets spanning 12 languages: Catalan, Czech, English, French, German, Hungarian, Lithuanian, Norwegian, Polish, Russian, Spanish, and Turkish. For the final evaluation, gold and predicted mentions are considered matching if they have overlapping head words, referred to as *head-match score*, and the CoNLL F1 head-match score is then macro-averaged over all 17 datasets.

For our submission, we adapt the model described in Toshniwal et al. (2020), which is based on the entity-ranking model originally proposed by Xia et al. (2020). We explore design decisions necessary to apply this English-based model to multilingual coreference resolution: data preprocessing steps, the pretrained language model encoder, and methods of joint training. Our best configuration outperforms the shared task baselines by a difference in head-match F1 score of +8.47, achieving an ultimate score of 65.43.

2 Related Work

Shared tasks have been instrumental in the development and evaluation of coreference resolution systems. Previous examples include CoNLL 2011 (Pradhan et al., 2011), CoNLL 2012 (Pradhan et al., 2012), and GAP (Webster et al., 2018, 2019). The CRAC 2023 shared task builds off the previous iteration, CRAC 2022 (Žabokrtský et al., 2022), with some modification of the datasets and evaluation procedure.

Entity-ranking models (Lee et al., 2017) of coreference resolution function by ranking a set of candidate entities to which each mention might refer. Xia et al. (2020) proposed a competitive neural entity-ranking model that processes mentions incrementally left-to-right. We analyze this method as implemented by Toshniwal et al. (2020). In contrast to existing work, we explore the potential of this model for multilingual generalization.

The best model of the previous CRAC 2022 shared task was that of Straka and Straková (2022), which consists of two stages: mention detection and coreference linking. The authors found that jointly training on multiple datasets led to better performance on the shared task than training several models, one per each individual dataset. The same finding was found in other submissions as well (Pražák and Konopik, 2022).

Existing analyses have considered the generalization of entity-ranking models across datasets, including when jointly trained on multiple datasets (Toshniwal et al., 2021; Xia and Van Durme, 2021; Porada et al., 2023). Although such work has focused on English-language coreference and not evaluated generalization to a multilingual collection of datasets. It is not clear, a priori, how well

Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution, pages 52–57
December 7, 2023

the constraints of an entity ranking model will generalize to phenomena not present in English coreference datasets such as zero anaphora.

3 Model

We evaluate the entity-ranking model implemented by Toshniwal et al. (2020). In this section, we first overview the model configuration and then outline the design decisions that we explore related to preprocessing, the pretrained encoder, and joint training. The high-level idea of the model is to first use a mention scorer to produce a set of mention candidates, then process the mentions left-to-right to determine if they refer to either a new or existing entity.

Configuration We start with the implementation and hyperparameters of Toshniwal et al. (2021). The model calculates coreference clusters for a document in the following way: first, embeddings are calculated for all spans of ≤ 20 subword tokens using a pretrained encoder. Each span embedding is scored using the mention scoring head described in Joshi et al. (2019), which is based on that originally proposed by Lee et al. (2017). This scoring head is trained with binary cross entropy loss to assign a positive score to annotated mentions and a negative score to all other spans. The top $0.4 \times \ell$ spans are considered as mention candidates and kept for the next step, where ℓ is the length of the document in terms of subword tokens. This set of mention candidates is further filtered by removing all spans with a negative score.

Then, the set of entities is initialized as $E = \{\}$ and the mention candidates are processed in a left-to-right order. When processed, each candidate m is scored against all entities $e \in E$ using a scoring function $s(m, e)$. If $\forall e \in E, s(m, e) < 0$ then m is added to the set E as a new entity. Otherwise, m is said to belong to the entity representation with the highest score $e^* = \text{argmax}_{e \in E}\ s(m, e)$ and the representation of e^* is updated to be the mean of all mention representations that the entity represents thus far. This method is referred to as the Unbounded Memory (U-MEM) model in the original work.

For training we use the default hyperparameters except for those that are specific to the pretrained encoder or number of training steps. We use the default optimizer of AdamW with a learning rate of 1e-5 for the pretrained encoder and 3e-4 for all other parameters.

Mention Heads The shared task evaluation requires the annotation of mention heads for each mention. We estimate mention heads from the provided dependency tree using heuristics provided by the Udapi library (Popel et al., 2017). Specifically, we use the command '`udapy -s corefud.MoveHead`'.

3.1 Preprocessing

We first convert the CoNLL-U files to a standardized JSON format using the file reader available in the Udapi Python library (Popel et al., 2017). We then tokenize each word independently using the pretrained encoder's tokenizer as implemented in Huggingface Transformers (Wolf et al., 2020). Finally, we concatenate all tokens together to produce a sequence of tokens representing the document.

Speaker Information We extract speaker information for each sentence from the sentence headers in the original CoNLL-U file. For example, the `CorefUD_English-GUM` corpora includes headers of the form "`# speaker = <SPEAKER_NAME>`" for certain documents. We include each speaker name s in the input at the beginning of the respective sentence. The name is formatted as "`<speaker>` s `</speaker>`" where `<speaker>` and `</speaker>` are randomly initialized tokens added to the model vocabulary. Including speakers as part of the text input such as in our approach was originally proposed by Wu et al. (2020).

Language Embedding We represent each language by a latent vector which is concatenated to the input of the entity-mention scoring function $s(m, e)$. The shared task datasets include 12 unique languages, so we define 12 such vectors. These language features are analogous to the OntoNotes genre features originally proposed by Wiseman et al. (2016).

Zero-anaphora When zeros appear in input (i.e., omitted pronouns that have been reconstructed in the coreference dataset), we represent these zeros as the underscore character '_' at training and test time since this is how they are represented in the CoNLL-U format.

3.2 Pretrained Encoder

We experimented with two pretrained encoders: XLM-RoBERTa (XLM-R; Conneau et al. 2020) and MT5 (Xue et al., 2021). To encode the document represented as a sequence of tokens, we split the sequence into chunks of maximum length L, encode the chunks using the pretrained encoder, and then concatenate the token encodings. Based on the sequence lengths the models were originally pretrained with, we use $L = 512$ for XLM-R and $L = 1024$ for MT5. We test using both the base and large model sizes for each encoder, up to 559M parameters for XLM-R and 995M parameters for MT5. In future work, it might be interesting to test RemBERT (Chung et al., 2021) as well, which was found by Straka and Straková (2022) to outperform XLM-R for multilingual coreference resolution.

3.3 Joint Training

We experiment with three methods for jointly training the model on all datasets: 1) **uniform weighting** where all datasets are sampled from equally; 2) **proportional weighting** where datasets are sampled proportional to the number of training examples in the dataset; and 3) **maximum weighting** where datasets are sampled from proportional to their training set size, except that training sets over some maximum threshold size are treated as if they are of that maximum size. This amounts to downscaling larger datasets to a maximum size. In our experiments we use 500 training examples as the maximum threshold.

4 Results

In this section we first present the results experimenting with each design decision, and then present the final submission performance. In preliminary experiments, we micro-average CoNLL F1 scores across all datasets for simplicity. For the final evaluation, CoNLL F1 scores are macro-averaged across datasets.

4.1 Pretrained Encoder

We experiment with both XLM-R and MT5 at the base and large model sizes. For these experiments, we report micro-averaged, exact-match CoNLL F1 on the development set (Table 1). We find that XLM-R, despite having fewer parameters and a shorter sequence length than MT5 outperforms the MT5 model. Possible explanations might be that: 1) MT5 was trained as an encoder-decoder model, while we use only the encoder for these experiments which creates a pretraining versus finetuning disparity that could hurt performance; or, 2) we finetuned the models with FP16 mixed precision whereas MT5 was pretrained with BF16 mixed precision.

Model		CoNLL F1
XLM-R	Base	71.9
	Large	**74.4**
MT5	Base	70.3
	Large	71.5

Table 1: Effects of the pre-trained encoder. CoNLL F1 score micro-averaged across all development sets.

Sample Weighting	CoNLL F1
Uniform	70.8
Proportional	71.9
Maximum	**72.9**

Table 2: Effects of the joint training method using the XLM-R base encoder. CoNLL F1 score micro-averaged across all development sets.

4.2 Joint Training

Next, we experiment with the three methods of joint training. For this experiment we use the XLM-R base encoder. We again evaluate using exact-match CoNLL F1 micro-averaged on the development set (Table 2). We find that the maximum weighting sampling method outperformed proportional sampling in this evaluation. For our final submission, we use a model first trained with proportional weighting for 50 epochs and next trained with maximum weighting for 50 epochs using early stopping on the development set.

4.3 Final Submission

Our final model achieves 65.43 F1 on the test set and fourth place in the competition (Table 3). We see a relatively high variance of the model ranking across languages (Table 4): for example, achieving second place on German-PotsdamCC and yet seventh place on both Czech-PDT and German-ParCorFull. This seems to be correlated with the relative size of the datasets, German-PotsdamCC being much larger than German-ParCorFull. Better performance on low-resource datasets is therefore a possible way to improve the performance of multilingual, entity-ranking models.

system	head-match	partial-match	exact-match	with singletons
1. CorPipe	74.90	73.33	71.46	76.82
2. Anonymous	70.41	69.23	67.09	73.20
3. Ondfa	69.19	68.93	53.01	68.37
4. **McGill**	65.43	64.56	63.13	68.23
5. DeepBlueAI	62.29	61.32	59.95	54.51
6. DFKI-Adapt	61.86	60.83	59.18	53.94
7. ITUNLP	59.53	58.49	56.89	52.07
8. BASELINE	56.96	56.28	54.75	49.32
9. DFKI-MPrompt	53.76	51.62	50.42	46.83

Table 3: Final F1 scores of all submissions. McGill (bolded) refers to our final submission which achieves fourth place in all categories except exact-match, for which it is in third place.

	ca	cs_1	cs_2	de_1	de_2	en_1	en_2	es	fr	hu	lt	pl	ru	hu	no_1	no_2	tr
Baseline	65.26	67.72	65.22	44.11	57.13	63.08	35.19	66.93	55.31	55.32	63.57	66.08	69.03	40.71	65.10	65.78	22.75
McGill	71.75	67.67	70.88	41.58	70.20	66.72	47.27	73.78	65.17	65.93	65.77	76.14	77.28	60.74	73.73	72.43	45.28
Δ	6.49	-0.05	5.66	-2.53	13.07	3.64	12.08	6.85	9.86	10.61	2.2	10.06	8.25	20.03	8.63	6.65	22.53

Table 4: Head-match CoNLL F1 scores of our final submission (McGill) as compared to the shared-task baseline for each language. $Delta$ is the difference in F1 score of both models. The datasets for each language, from left to right, are: ca_ancora, cs_pcedt, cs_pdt, de_parcorfull, de_potsdamcc, en_gum, en_parcorfull, es_ancora, fr_democrat, hu_szegedkoref, lt_lcc, pl_pcc, ru_rucor, hu_korkor, no_bokmaalnarc, no_nynorsknarc, and tr_itcc.

5 Conclusion

We adapt an entity-ranking coreference resolution model to multilingual coreference resolution for the CRAC 2023 shared task. We explore the method of training and joint encoder, finally using XLM-R large and a rescaled dataset weighting in our submission. This method achieved fourth place of nine submissions in the shared task.

Acknowledgements

The authors acknowledge the material support of NVIDIA in the form of computational resources. Ian Porada is supported by a fellowship from the Fonds de recherche du Québec (FRQ). Jackie Chi Kit Cheung is supported by the Canada CIFAR AI Chair program.

References

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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.

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.

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.

Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman, Anna Nedoluzhko, Kutay Acar, Peter Bourgonje, Silvie Cinková, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, Pauline Krielke, Frédéric Landragin, Ekaterina Lapshinova-Koltunski, Petter Mæhlum, M.Antònia Martí, Marie Mikulová, Anders Nøklestad, Maciej Ogrodniczuk, Lilja Øvrelid, Tuğba Pamay Arslan, Marta Recasens, Per Erik Solberg, Manfred Stede, Milan Straka, Svetlana Toldova, Noémi Vadász, Erik Velldal, Veronika Vincze, Amir Zeldes, and Voldemaras Žitkus. 2022. Coreference in universal dependencies 1.1 (CorefUD 1.1).

LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2023. Investigating failures to generalize for coreference resolution models.

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 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, Portland, Oregon, USA. 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.

Milan Straka and Jana Straková. 2022. ÚFAL CorPipe at CRAC 2022: Effectivity of multilingual models for coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 28–37, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. 2020. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online. Association for Computational Linguistics.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kellie Webster, Marta R. Costa-jussà, Christian Hardmeier, and Will Radford, editors. 2019. *GAP Shared Task Overview*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, California. Association for Computational Linguistics.

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.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, 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.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.